Parcours Data Scientist

SOUTENANCE PROJET 3 : CONCEVEZ UNE APPLICATION AU SERVICE DE LA SANTÉ PUBLIQUE

22 avril 2022

Santé publique France

GAUTHIER RAULT
PARCOURS DATA SCIENTIST
CHEZ OPENCLASSROOMS

EVALUATEUR MADAME FATOU SALL

Daniel KEYS MORAN (DKM)
“You can have data without information, but you cannot have information without data.”

"Vous pouvez avoir des données sans informations, mais vous ne pouvez pas avoir des informations sans données."

ORDRE DU JOUR

PRÉSENTATION (20MIN)

Rappel de l’appel à projets et explication de l’idée d’application (2 mn)

Démarche méthodologique de nettoyage (8 mn)

Démarche méthodologique d’exploration de données (8 mn)

En synthèse, présentation des faits pertinents pour l’application (2 mn)

DISCUSSION (5MIN)

DEBRIEFING (5MIN)

CONTEXTE

Santé publique France
a lancé un appel à projets pour trouver des idées
innovantes d'applications en lien avec l'alimentation

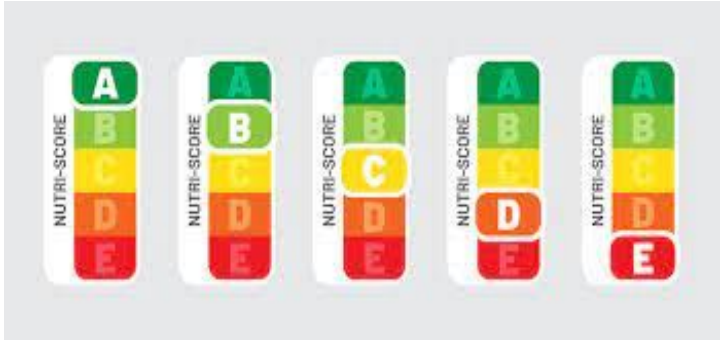
En s'appuyant sur les données d'Open Food Facts, nous répondrons à la question:

Quel type application peut aider le grand public à améliorer son alimentation?

OPEN FOOD FACTS

IDÉE D'APPLICATION

L'OBJECTIF DE L'APPLICATION EST DE PROPOSER, POUR LA FRANCE, UN PRODUIT SIMILAIRE AVEC UN MEILLEUR NUTRI-SCORE



Source :

https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/parcours-data-scientist/p2/fr.openfoodfacts.org.products.csv.zip

PHASE D'INSPECTION DU JEU DE DONNÉES À DISPOSITION

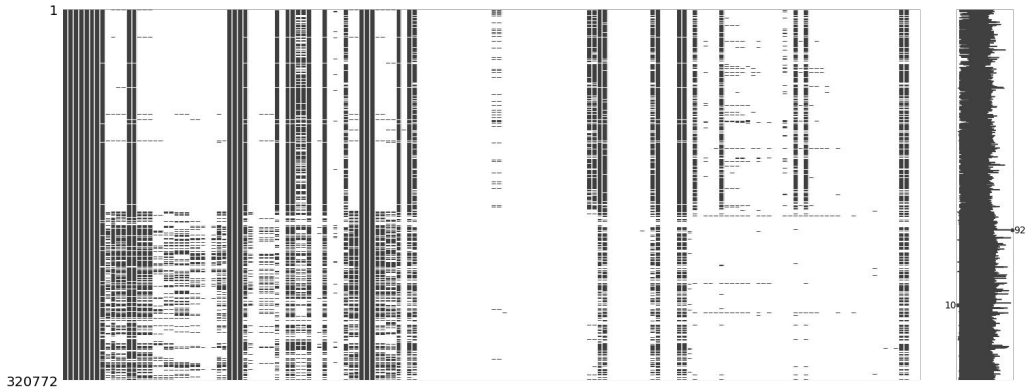
QUANTITATIF

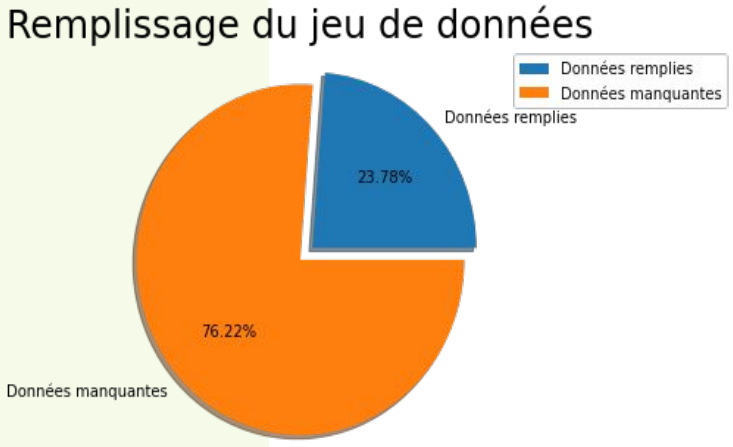
```
: Food.info()
```

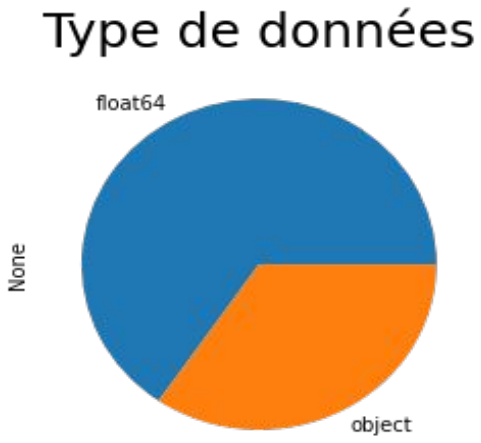
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 320772 entries, 0 to 320771
Columns: 162 entries, code to water-hardness_100g
dtypes: float64(106), object(56)
```

PHASE D'INSPECTION DU JEU DE DONNÉES À DISPOSITION

QUALITATIF







DÉMARCHE MÉTHODOLOGIQUE DE NETTOYAGE

1ÈRE ÉTAPE

Limiter le nombre de NaN trop conséquent
Suppression des colonnes ayant moins de 50% de remplissage.
Sans utiliser de fonction car sinon dès l'ajout d'une nouvelle entrée, elle serait retirée instantanément)

```
Food = Food[Food.columns[Food.isna().sum()/Food.shape[0]<0.5]]
```

2ÈME ÉTAPE

Se concentrer sur la France exclusivement

```
Food = Food.loc[Food["countries_fr"]== "France"]
```

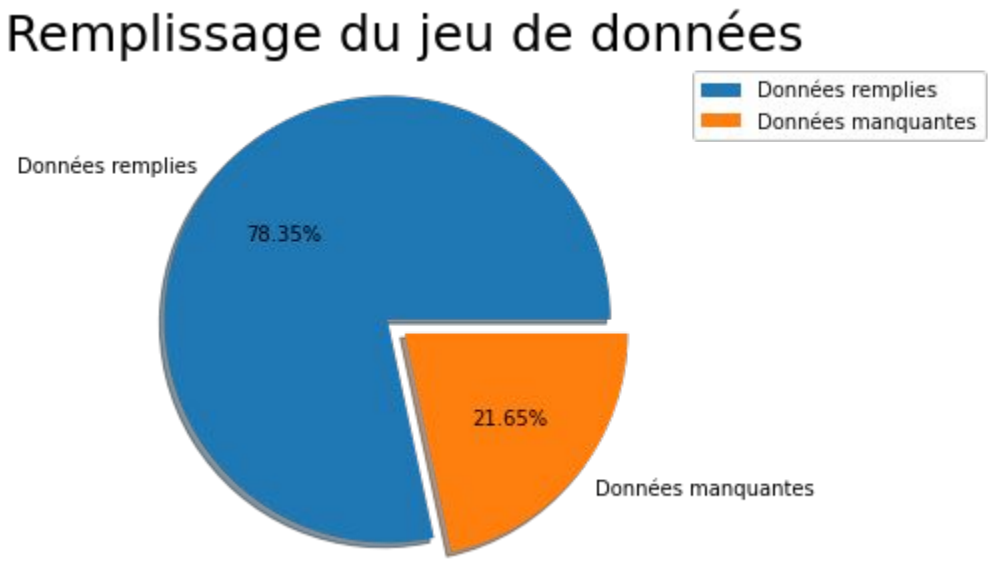
3ÈME ÉTAPE

Les produits n'ayant pas de nom sont amputés

```
Food = Food.loc[~Food["product_name"].isna()]
```

DÉMARCHE MÉTHODOLOGIQUE DE NETTOYAGE

Ces actions inversent le nombre de données remplies et manquantes



DÉMARCHE MÉTHODOLOGIQUE DE NETTOYAGE

L'objectif de l'application est de proposer, pour la France, un produit similaire avec un meilleur nutri-score

Pour rappel :https://fr.wikipedia.org/wiki/nutri-score

Les variables n'ayant pas d'intérêts pour l'application sont écartées

```
Food = Food.drop(columns=["created_datetime",
                          "states_fr",
                          "created_t",
                          "creator",
                          "last_modified_datetime",
                          "ingredients_that_may_be_from_palm_oil_n",
                          "last_modified_t",
                          "countries",
                          "countries_tags",
                          "serving_size",
                          "states",
                          "states_tags",
                          "brands_tags",
                          "nutrition-score-uk_100g"])
```

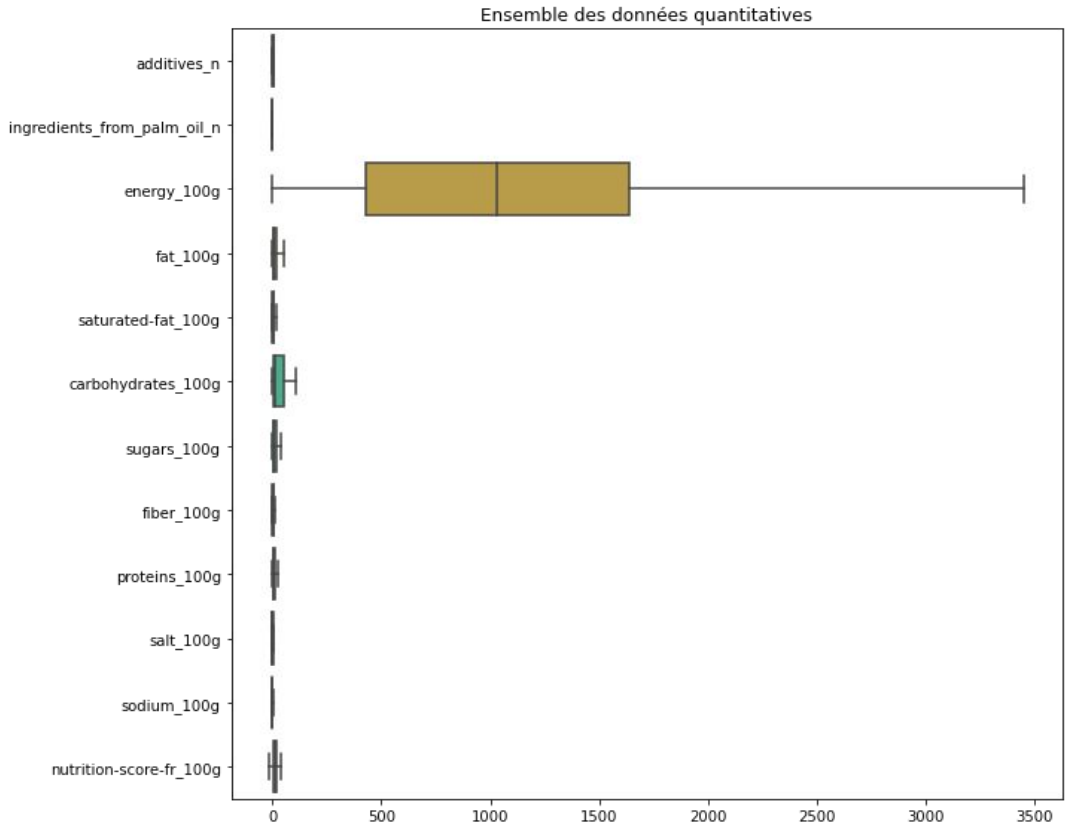
DÉMARCHE MÉTHODOLOGIQUE DE NETTOYAGE

Mise en évidence que la colonne code contient des doublons.
Les doublons sont retirés pour garder la première occurrence.

```
Food.drop_duplicates(subset ="code", keep = 'first', inplace=True)
```

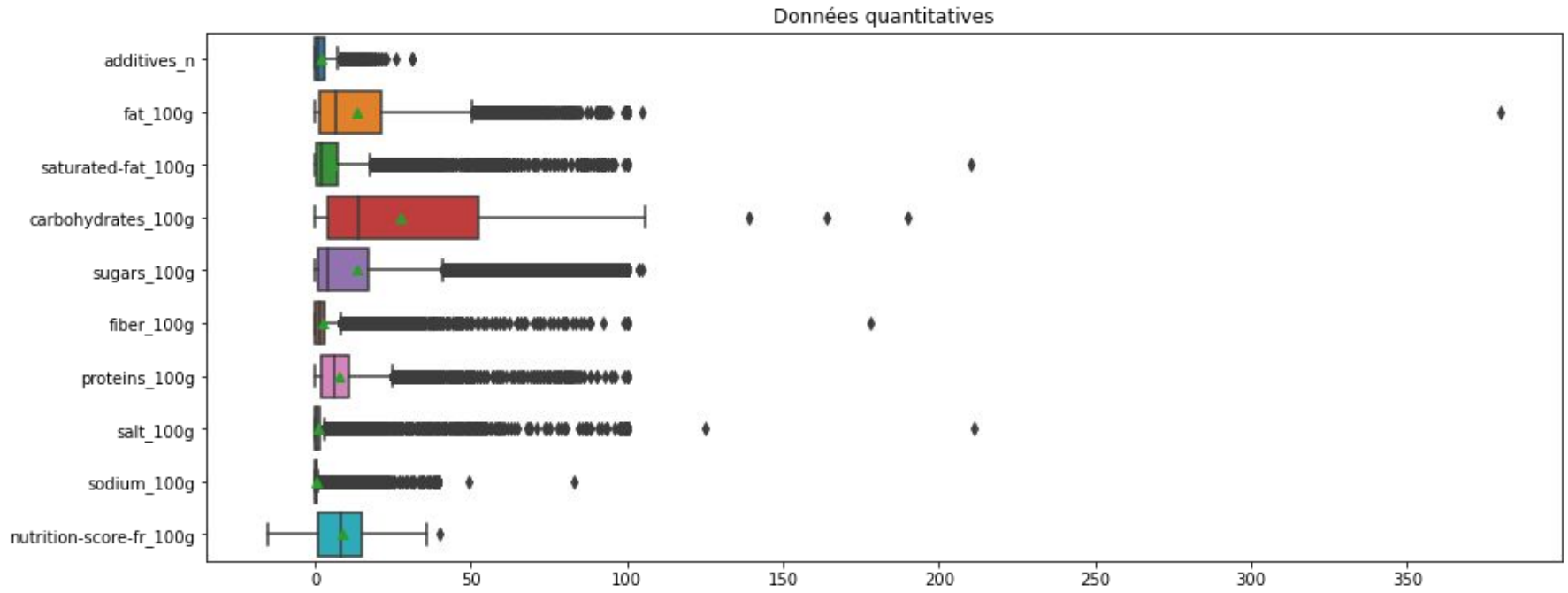
DÉMARCHE MÉTHODOLOGIQUE DE NETTOYAGE

- 'energy_100g' est exprimée en kilojoules (kj) donc la base de 100g maximum n'est pas cohérente pour cette feature.
- Une recherche internet montre que la valeur maximum est 3500kj.
- Une fonction vérifie valeur >3500 pour energy et pour les autres features (excluant nutriscore) >100g et <0g.
- Nutriscore valeurs comprises entre - 15 et 40
- Les valeurs détectées sont converties en None pour un retraitement futur.



DÉMARCHE MÉTHODOLOGIQUE DE NETTOYAGE

Observation des différents outliers avec des valeurs aberrantes



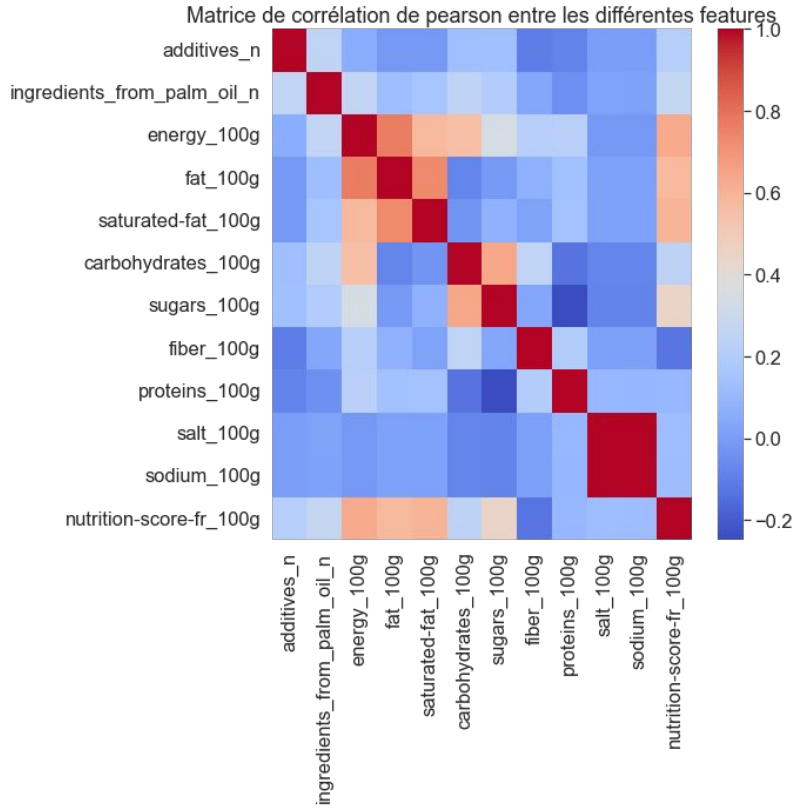
DÉMARCHE MÉTHODOLOGIQUE DE NETTOYAGE

- Utilisation de la fonction iterativeimputer de la librairie scikit-learn.

Pour l'utiliser, il y a un prérequis sur la corrélation entre les variables utilisées.

- Modélisation d'une heatmap retenant les variables ayant un coefficient de corrélation >0.5

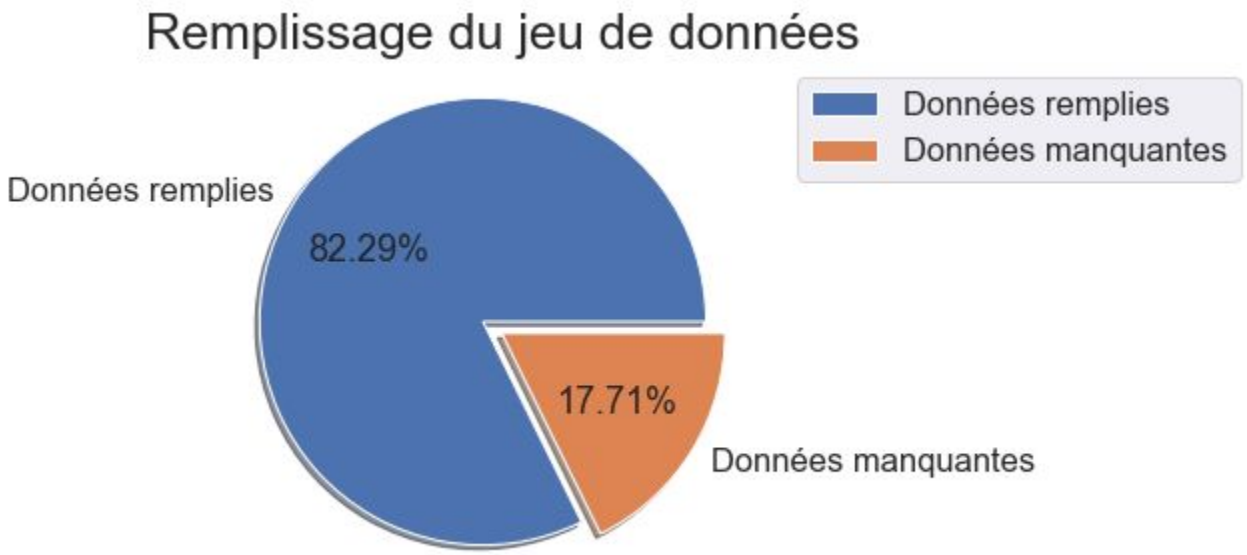
```
# limite de corrélation > 0.5
List_energy_fat = Food[["energy_100g","fat_100g"]].columns
List_energy_saturated_fat = Food[["energy_100g","saturated-fat_100g"]].columns
List_energy_carbohydrates = Food[["energy_100g","carbohydrates_100g"]].columns
List_energy_nutrition_score_fr = Food[["energy_100g","nutrition-score-fr_100g"]].columns
List_sugars_carbohydrates = Food[["sugars_100g","carbohydrates_100g"]].columns

List_fat_saturated_fat = Food[["fat_100g","saturated-fat_100g"]].columns
List_fat_nutrition_score_fr = Food[["fat_100g","nutrition-score-fr_100g"]].columns
List_saturated_fat_nutrition_score_fr = Food[["saturated-fat_100g","nutrition-score-fr_100g"]].columns
```



DÉMARCHE MÉTHODOLOGIQUE DE NETTOYAGE

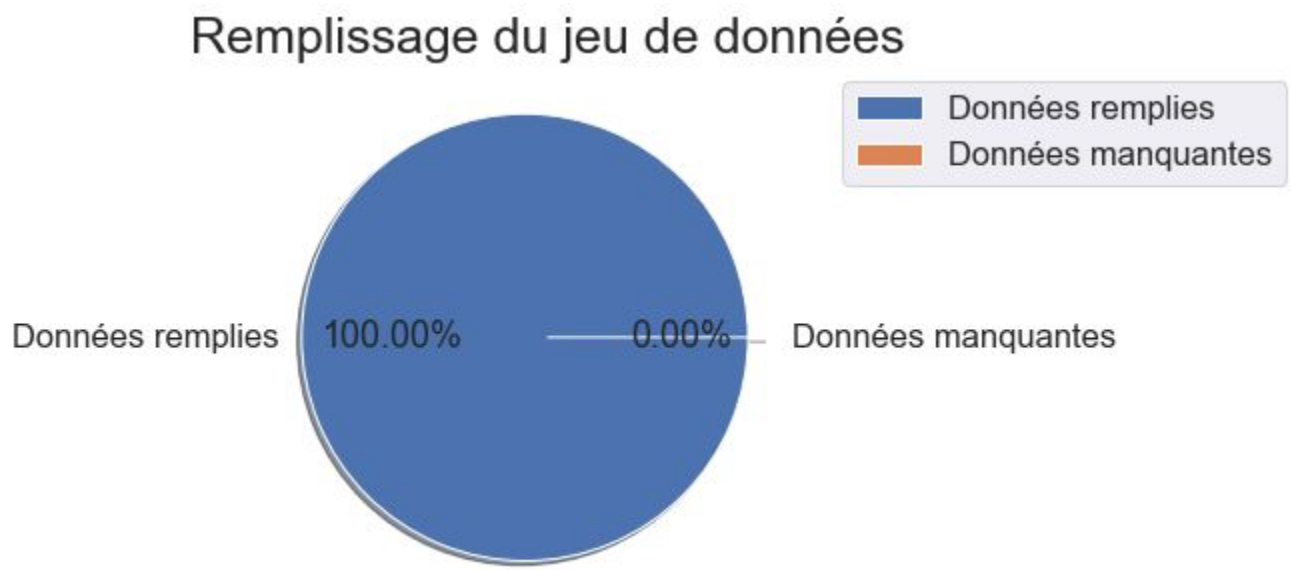
- La fonction iterativeimputer a permis d'augmenter de plus de 10 points de pourcentage les données remplies.

Méthode efficace mais avec le travers d'alimenter le dataset avec des valeurs aberrantes.

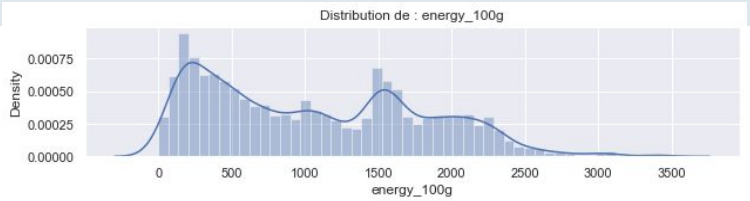


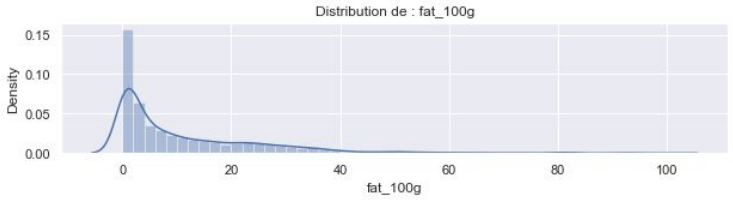
DÉMARCHE MÉTHODOLOGIQUE DE NETTOYAGE

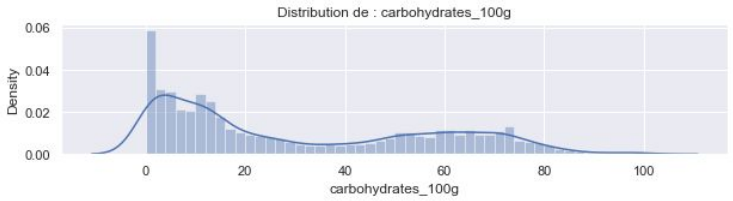
- Finalisation de la démarche de nettoyage.
- Application d'un dropna sur les valeurs manquantes afin d'avoir un dataset complètement rempli.

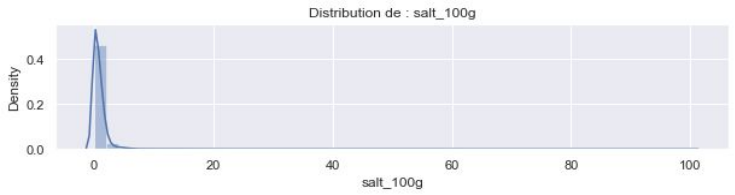


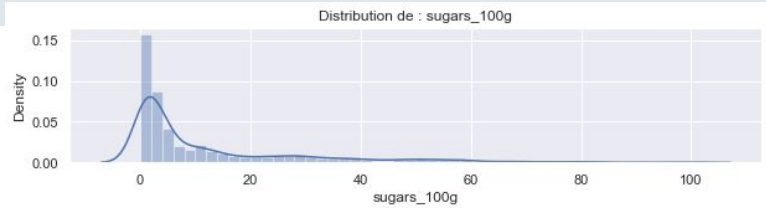
DÉMARCHE MÉTHODOLOGIQUE D'EXPLORATION DE DONNÉES

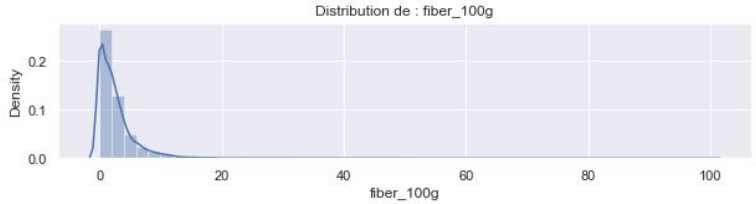


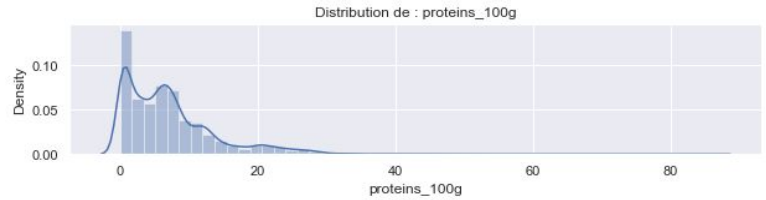


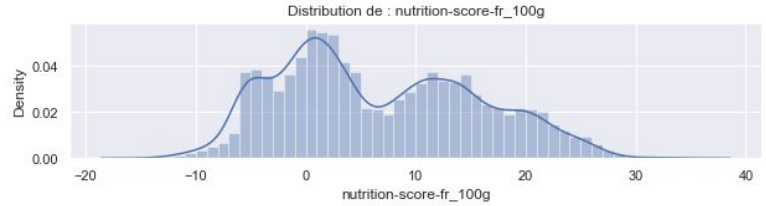




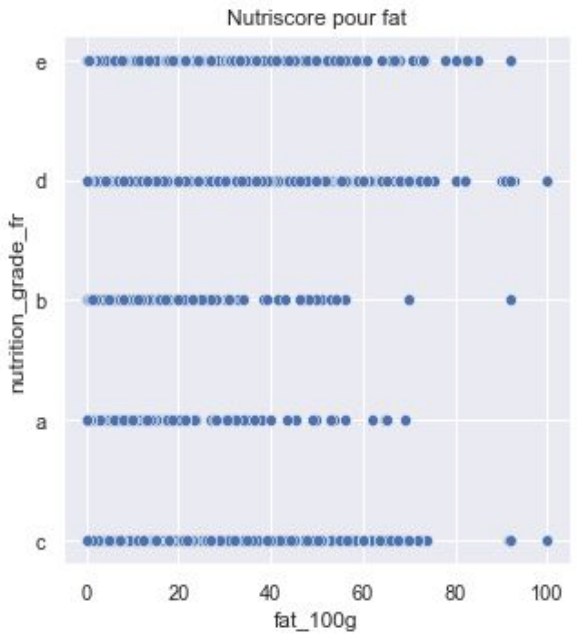


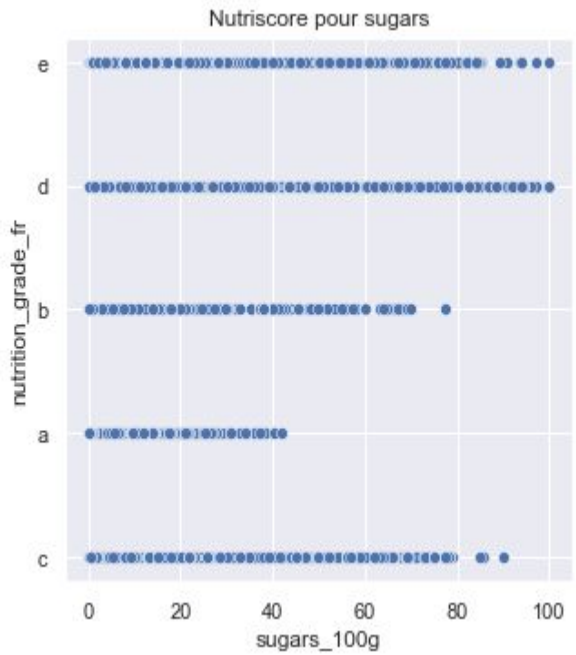


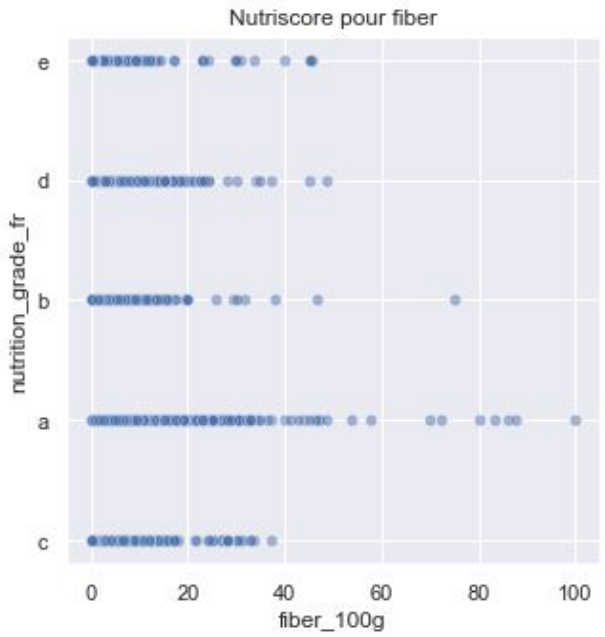




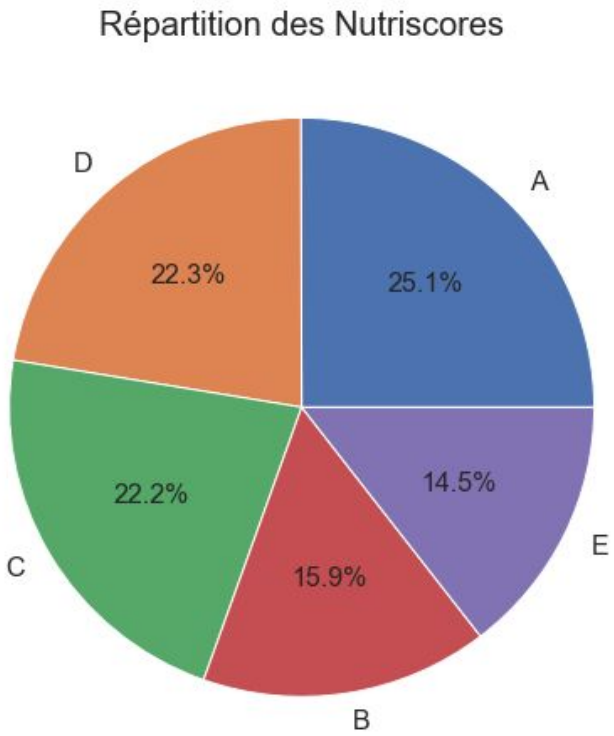
DÉMARCHE MÉTHODOLOGIQUE D'EXPLORATION DE DONNÉES





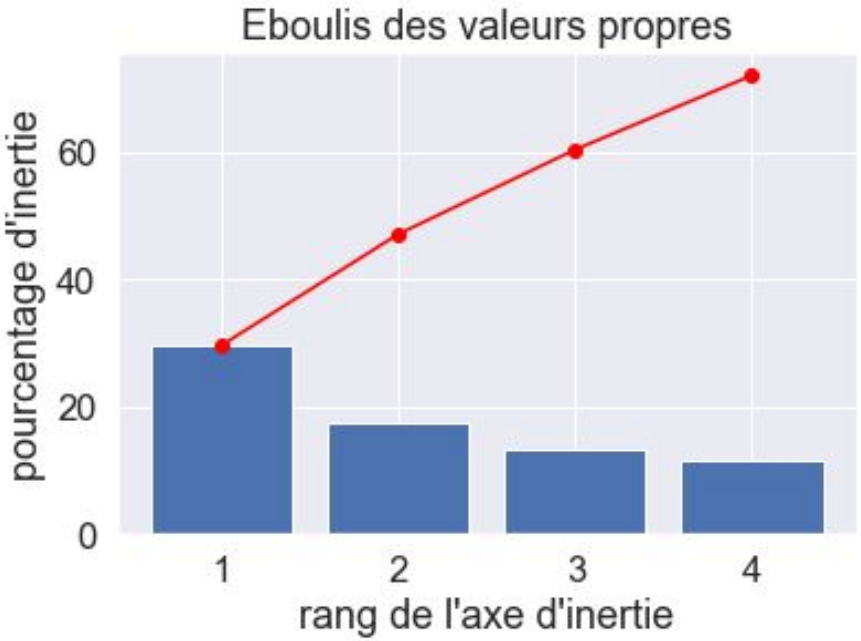


DÉMARCHE MÉTHODOLOGIQUE D'EXPLORATION DE DONNÉES

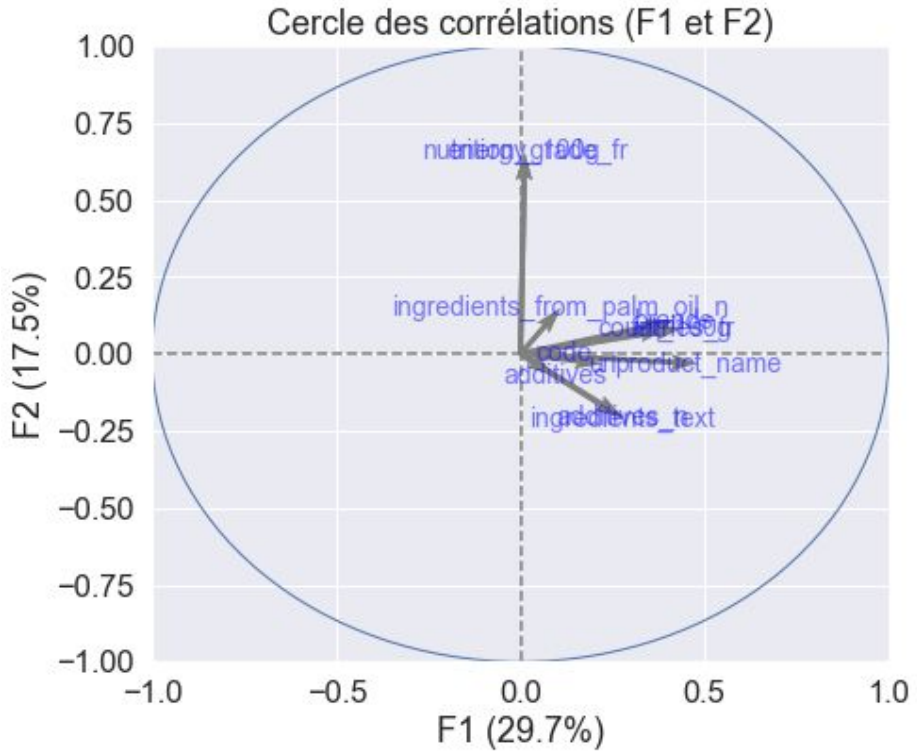


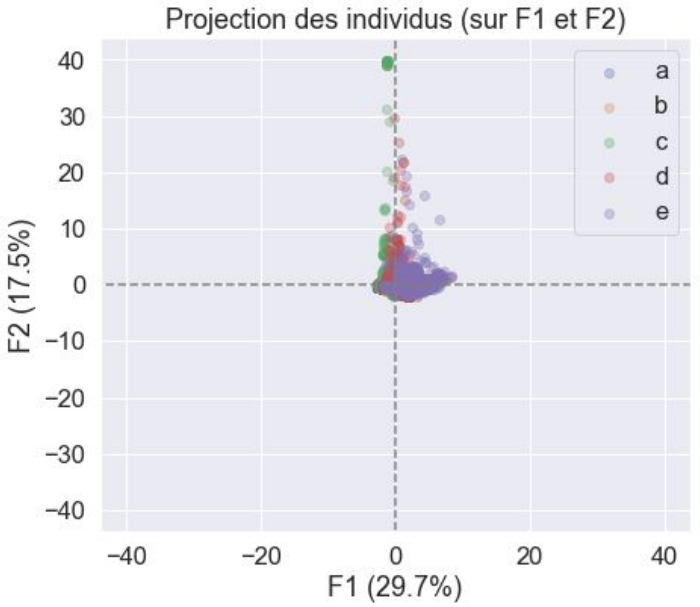
DÉMARCHE MÉTHODOLOGIQUE D'EXPLORATION DE DONNÉES

ACP/PCA
Analyse en composantes principales/
Principal component analysis

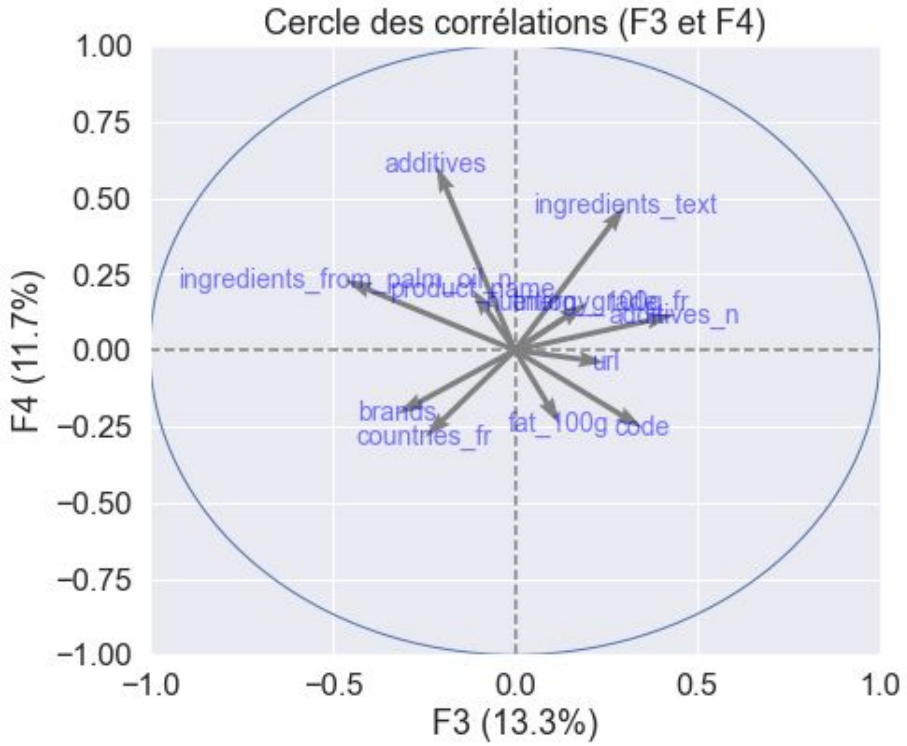


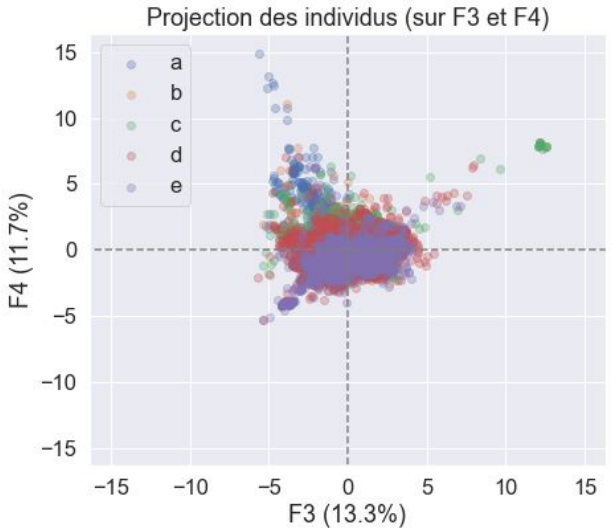
DÉMARCHE MÉTHODOLOGIQUE D'EXPLORATION DE DONNÉES





DÉMARCHE MÉTHODOLOGIQUE D'EXPLORATION DE DONNÉES





DÉMARCHE MÉTHODOLOGIQUE D'EXPLORATION DE DONNÉES

ANOVA
Analysis of variance

```
: X = Food['nutrition_grade_fr'] # qualitative
  Y = Food['fat_100g'] # quantitative

  eta_squared(X,Y)
```

```
0.31302844683535097
```

```
: Y = Food['energy_100g']

  eta_squared(X,Y)
```

```
0.30650849563200305
```

```
: Y = Food['saturated-fat_100g']
  eta_squared(X,Y)
```

```
0.3719535231469069
```

```
: Y = Food['carbohydrates_100g']
  eta_squared(X,Y)
```

```
0.0669800147472224
```

```
: Y = Food['sugars_100g']
  eta_squared(X,Y)
```

```
0.24111134369228168
```

```
: Y = Food['fiber_100g']
  eta_squared(X,Y)
```

```
0.03734590014756734
```

```
: Y = Food['proteins_100g']
  eta_squared(X,Y)
```

```
0.01998887430119569
```

```
: Y = Food['salt_100g']
  eta_squared(X,Y)
```

```
0.014532130449342293
```

PRÉSENTATION DES FAITS PERTINENTS POUR L'APPLICATION

FOODAGA

NUTRI-SCORE

A B C D E

NUTRI-SCORE

A B C D E

PRÉSENTATION DES FAITS PERTINENTS POUR L'APPLICATION

FOODAGA

Au travers de l'application les utilisateurs pourront améliorer leur alimentation :

- en sucre, en graisses, en calories qui provoquent des maladies
- les fibres qui sont bonnes pour la santé

Ce qui a pour conséquence de minimiser les dépenses de santé car la population est en meilleure santé

CONCLUSION SUR LE JEU DE DONNÉES

Recommandations :

- Obtenir d'autres jeux de données pour augmenter la complétude et la crédibilité
- Croiser les résultats avec ces autres jeux de données

Diversité des données
Données intéressantes récoltées

Faible complétude
Variables non décrites
Rempli par des utilisateurs

OPEN FOOD FACTS

Yuka

Alim'confiance

Etiquettable

OUVERTURE SUR D'AUTRES UNIVERS

Autres facteurs pouvant être pris en compte autres que le nutriscore :

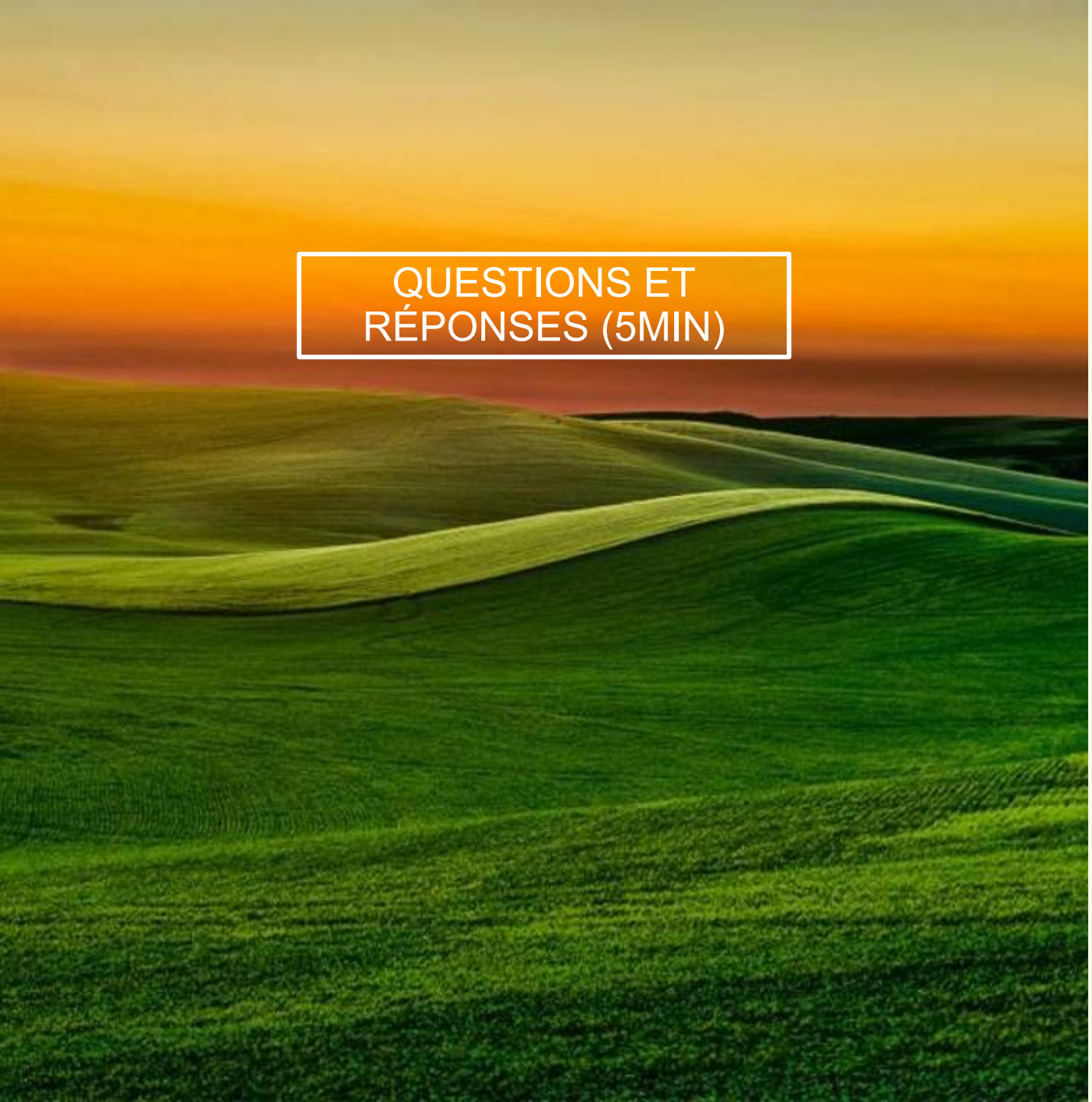
Une consommation engagée (de proximité, environnementale, politique, vegan)

Une alimentation spécifique et personnalisée (si je suis diabétique, me proposer des recommandations adaptées avec moins de sucre)

Pour bébé, sportif

AXES D’AMÉLIORATIONS DE L’ ÉTUDE

- Travailler d’avantage sur la complétude et fraîcheur des données
- Vérifier que les valeurs 100g en les additionnant respecte bien 100g
- Utiliser d’autres algorithmes pour palier les NaN (KNN)
- Conserver toutes les features pour augmenter les possibilités
- Faire une ACM, Analyse des correspondances multiples
- Améliorer les visualisations et en trouver d’autres
- Améliorer le code python et le simplifier


QUESTIONS ET
RÉPONSES (5MIN)

MERCI POUR
VOTRE ATTENTION