



# Projet3 Concevez une application au service de la santé publique

CONCEVOIR UNE APPLICATION POUR LES SPORTIFS

AL SAMMAN Wassim – Data Scientist Apprenti

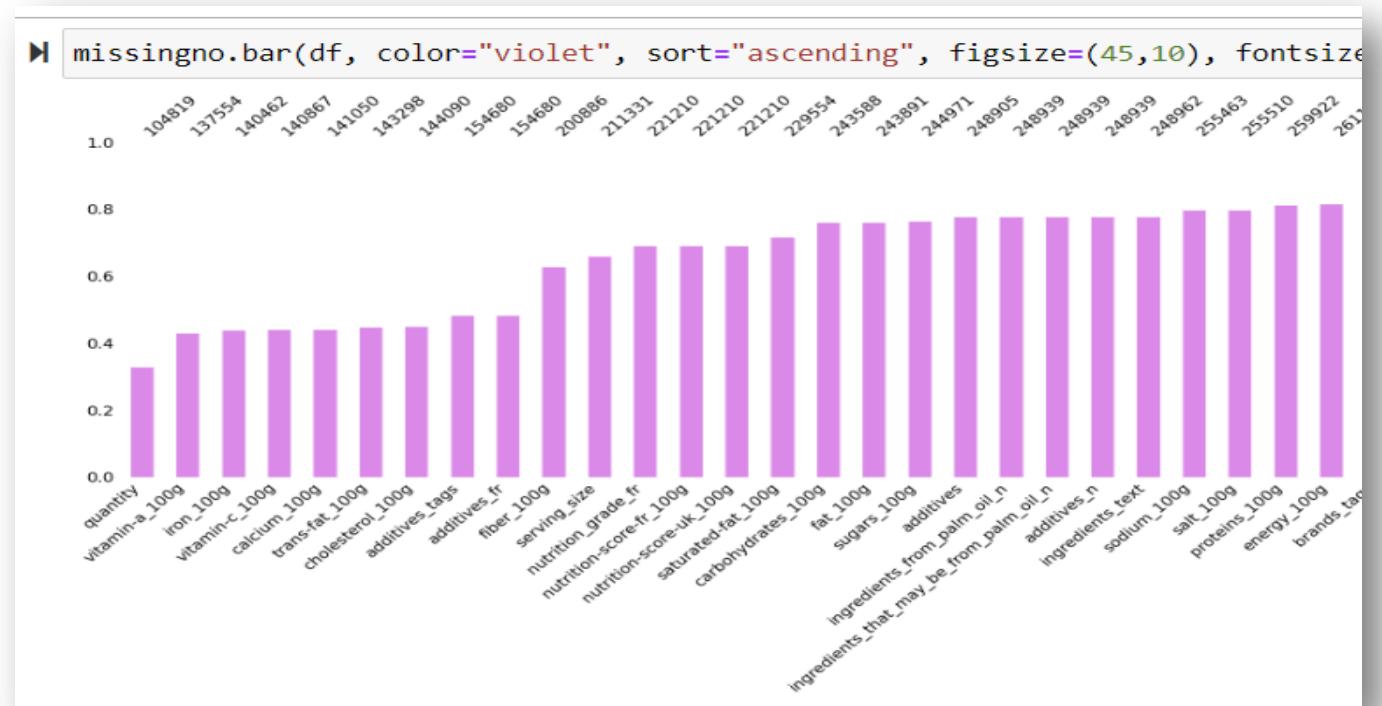
Mentor : Panayotis PAPOUTSIS

# Introduction

- ▶ Le projet de l'agence Santé Publique France.
- ▶ Idée d'une application innovante.
- ▶ Mon idée de l'application :
  - ▶ Créer une application pour les sportifs.
  - ▶ Aider à choisir l'alimentation la plus adaptée.
  - ▶ Donner le pourcentage de protéine et d'autres aliments importants.
  - ▶ Aider les utilisateurs pour améliorer leur choix d'alimentation.

# Préparer et nettoyer les données

- ▶ Supprimer les colonnes avec un grand nombre de données manquantes.
- ▶ Premier choix des colonnes:
  - ▶ **Les colonnes:**  
code, url, product\_name,  
quantity, ingredients\_text.
  - ▶ **Les colonnes:**  
proteins\_100g,  
carbohydrates\_100g.
  - ▶ **La colonne:** nutrition\_grade\_fr
  - ▶ **D'autres colonnes**

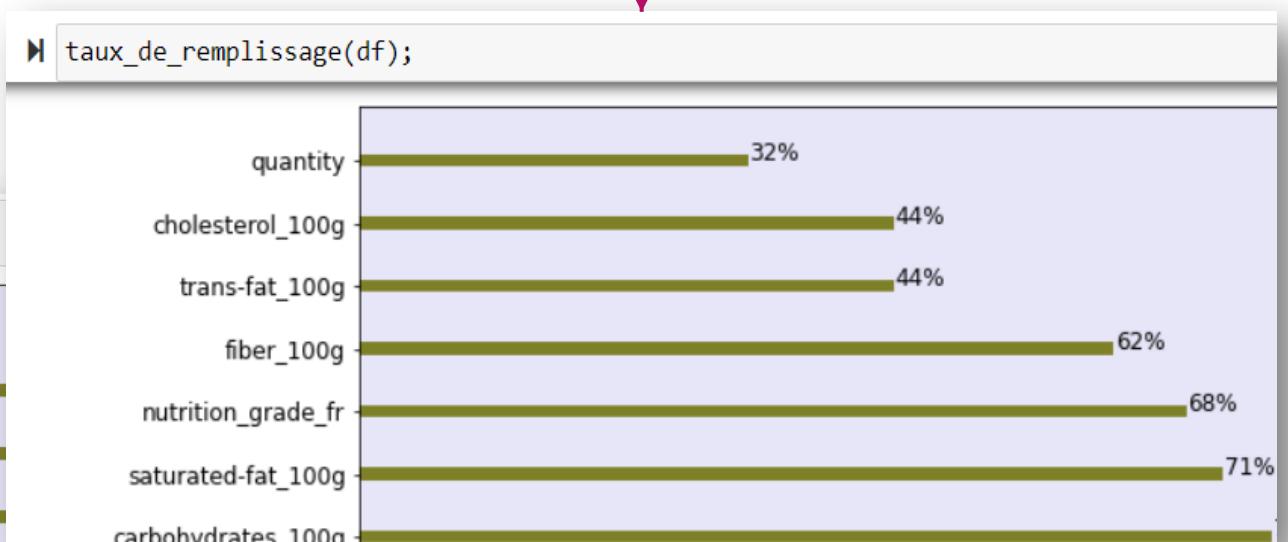


# Préparer et nettoyer les données

- ▶ Vérifier le taux de remplissage des colonnes choisies.
- ▶ Rechoisir les colonnes.
- ▶ Traiter les valeurs aberrantes.
- ▶ Revérifier le taux de remplissage.

```
▶ df=VarAberr0100(df)
  taux_de_remplissage(df);

  energy_100g      5%
  nutrition_grade_fr
  saturated-fat_100g
  carbohydrates_100g
```



# Préparer et nettoyer les données

## ► Remplir les valeurs manquantes:

- ▶ Une fonction qui réaliser tout le remplissage.
- ▶ Supprimer les lignes pour les valeurs manquantes de deux premières colonnes (code, url).
- ▶ Remplir les colonnes qui contient du text par la valeur la plus répétitive avec méthode fillna().
- ▶ Remplir par moyenne les colonnes numériques qui se termines par '100g'.
- ▶ Remplir la colonne catégorielle nutrition\_grade\_fr avec la méthode de Knn.

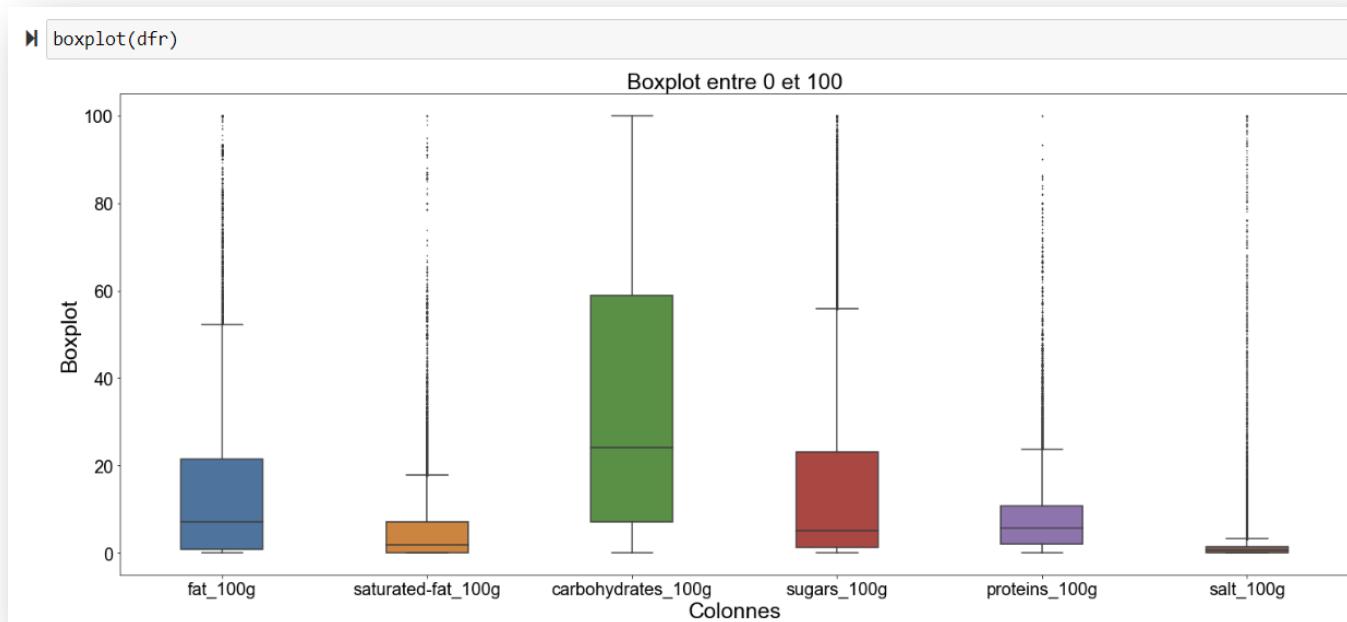
- ▶ Exporter le DataFrame final et rempli pour le notebook d'exploration.

```
dfr=remplir(df)
dfr.isnull().sum()
```

```
:5: code          0
    url          0
    product_name 0
    ingredients_text 0
    nutrition_grade_fr 0
    fat_100g      0
    saturated-fat_100g 0
    carbohydrates_100g 0
    sugars_100g    0
    proteins_100g 0
    salt_100g      0
dtype: int64
```

# Exploration des données\_Analyse Uni-Variée

- ▶ Utiliser la fonction describe.
- ▶ La méthode de boxplot.

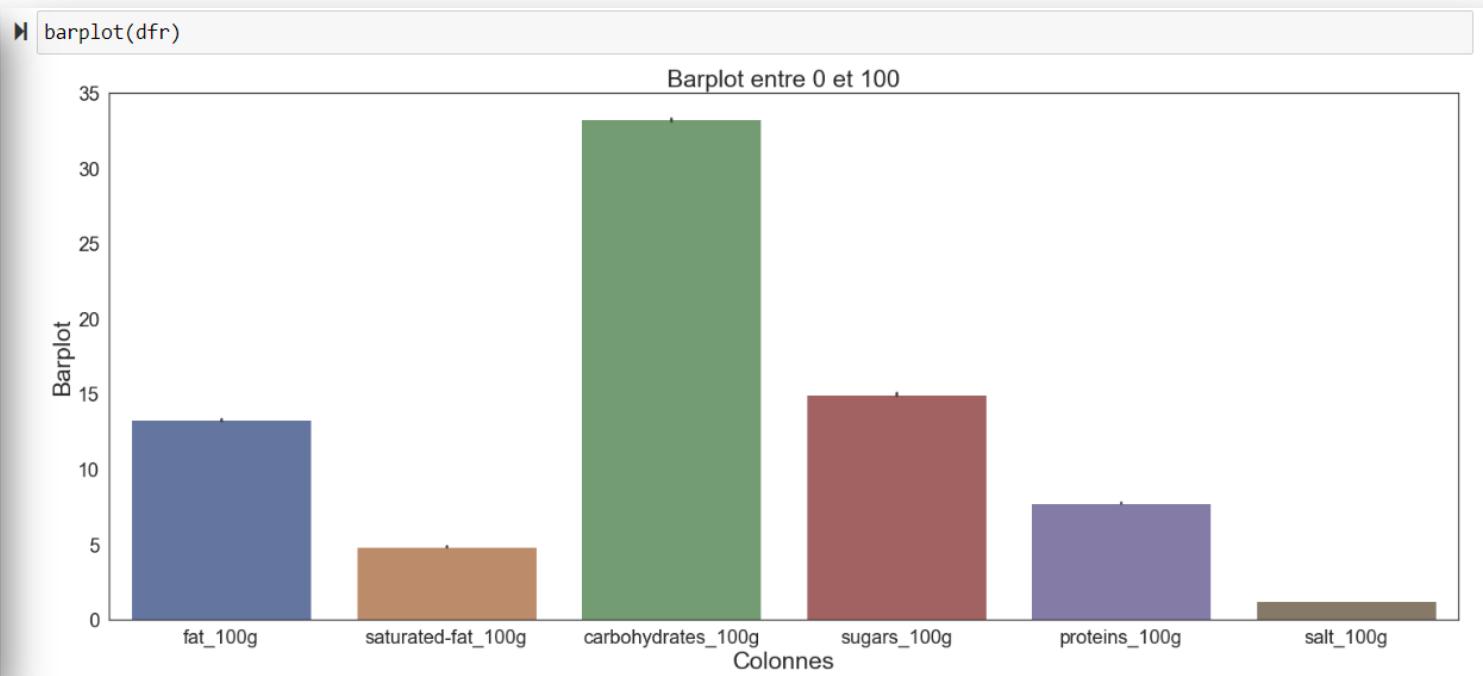
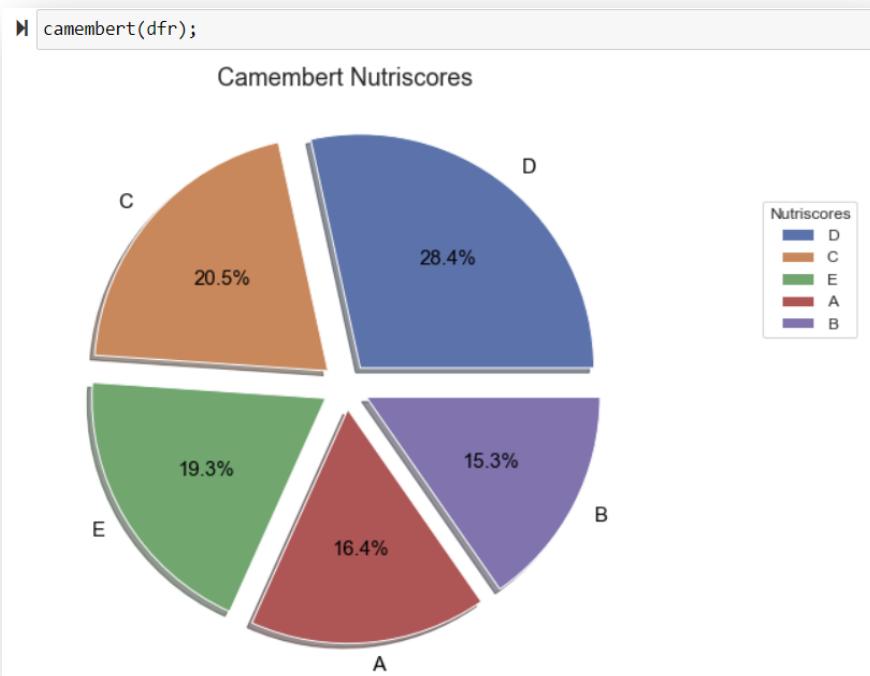


dfr.describe()

	fat_100g	saturated-fat_100g
count	201236.000000	201236.000000
mean	13.303567	4.8742
std	16.067705	7.3119
min	0.000000	0.0000
25%	0.830000	0.0000
50%	7.140000	1.7900
75%	21.430000	7.1400
max	100.000000	100.0000

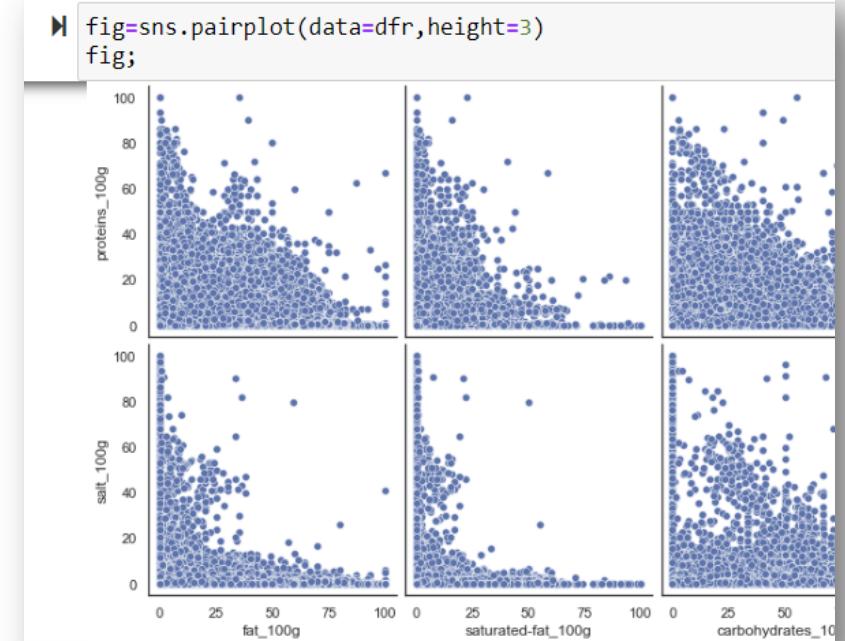
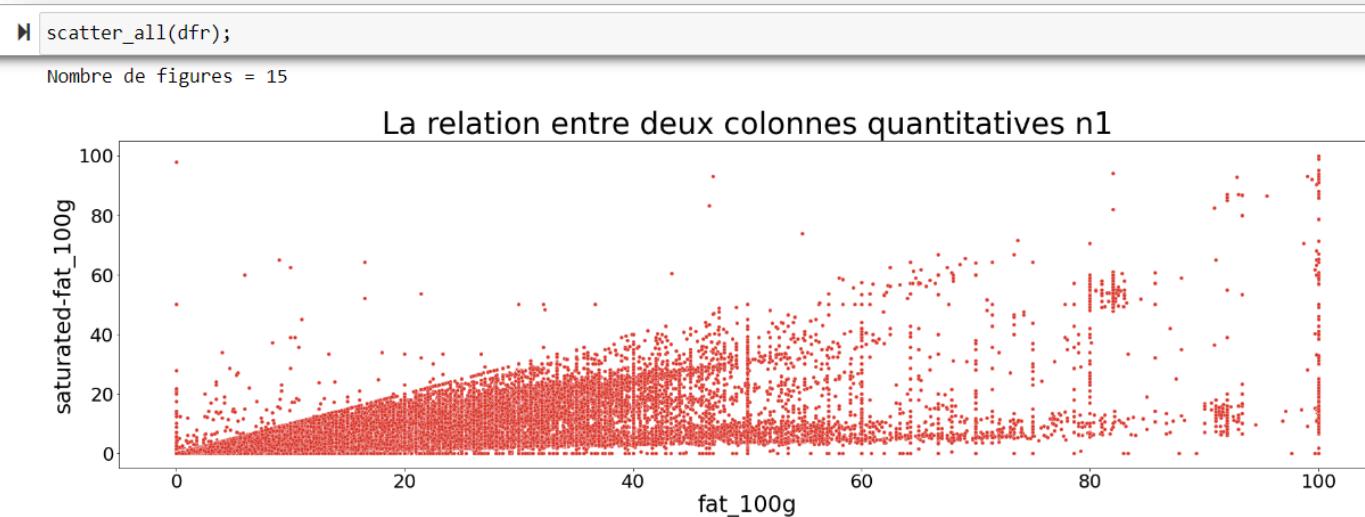
# Exploration des données\_Analyse Uni-Variée

- ▶ La méthode de barplot.
- ▶ Camembert.



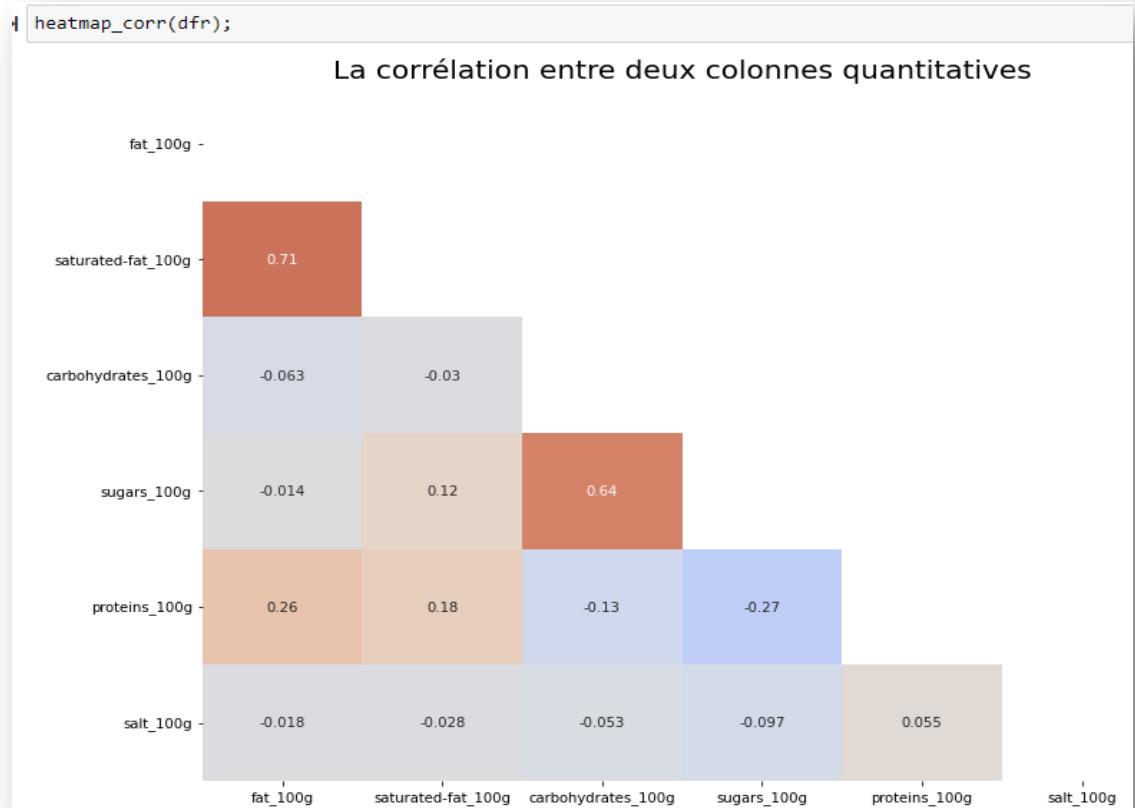
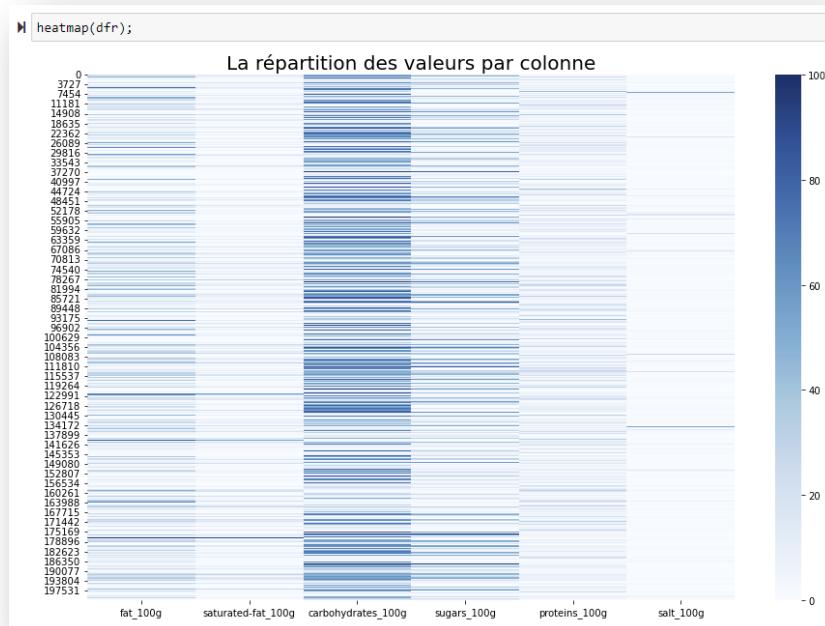
# Exploration des données\_Analyse Bi-Variée

- ▶ Scatterplot, la fonction scatter\_all.
- ▶ Pairplot.



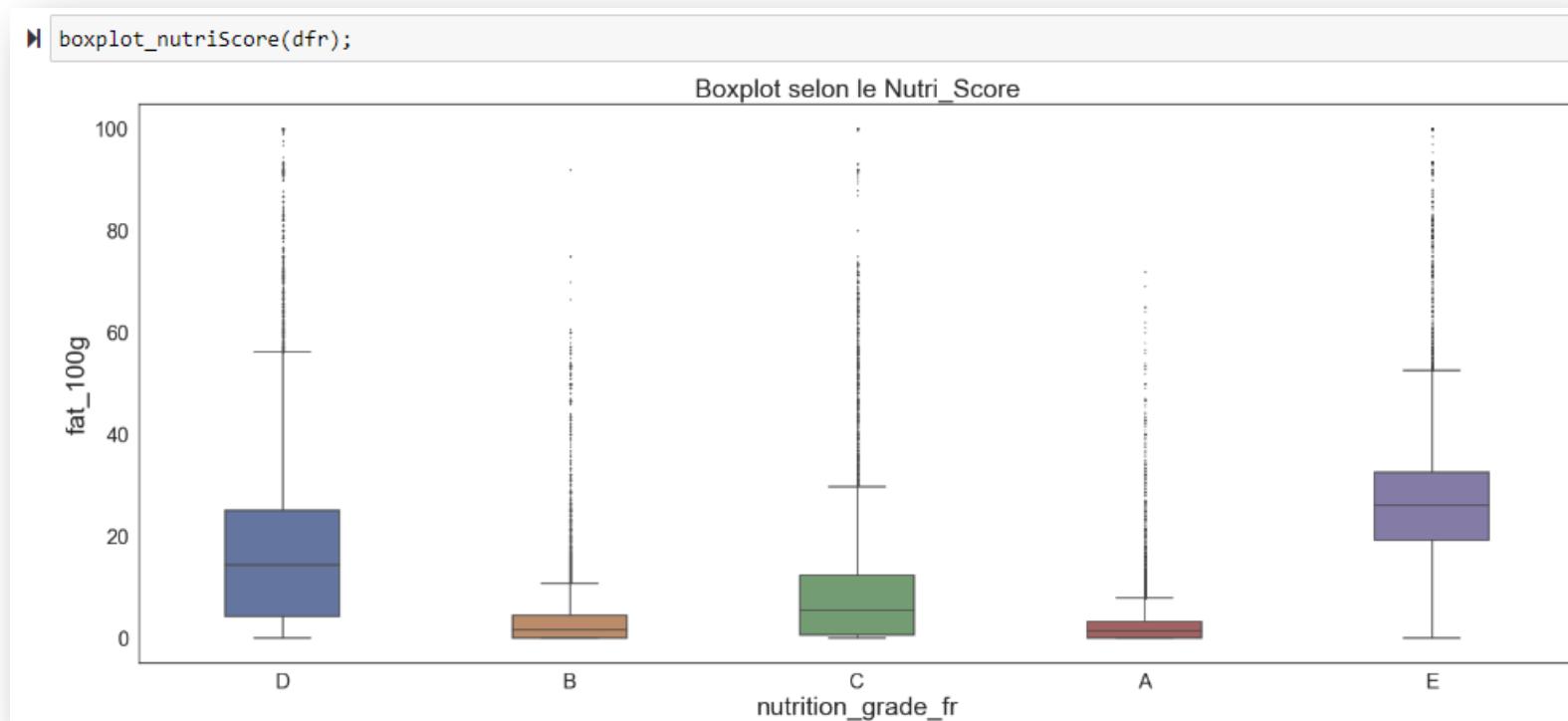
# Exploration des données\_Analyse Bi-Variée

- ▶ Heatmap.
- ▶ Heatmap de corrélation.



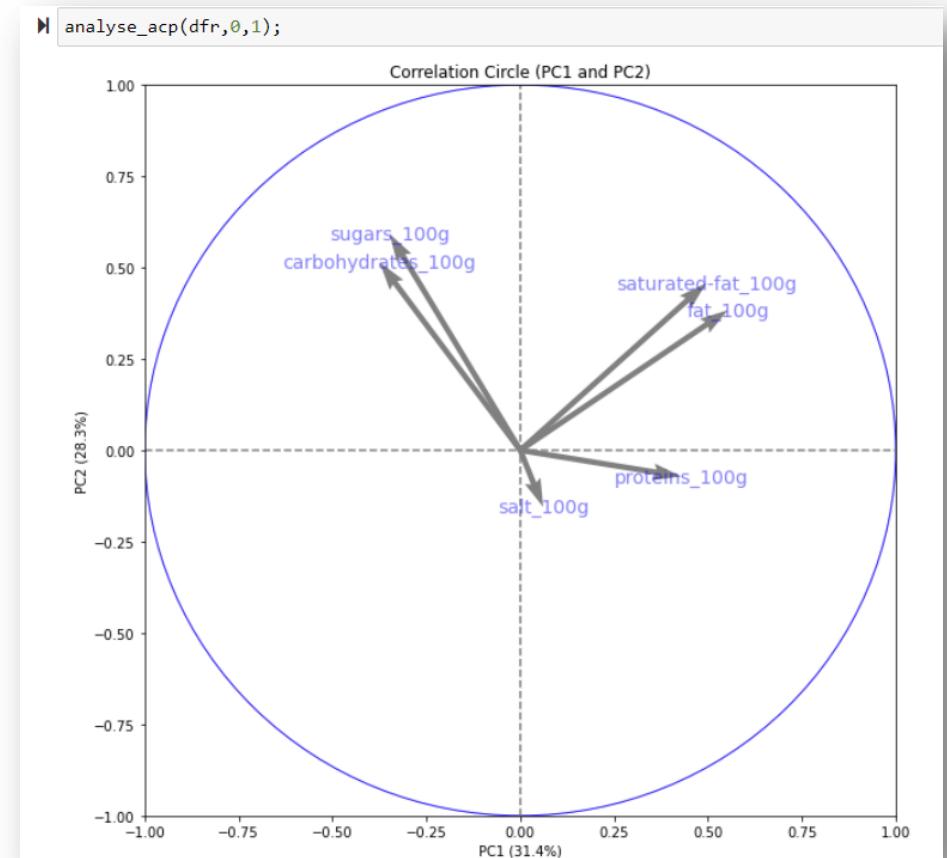
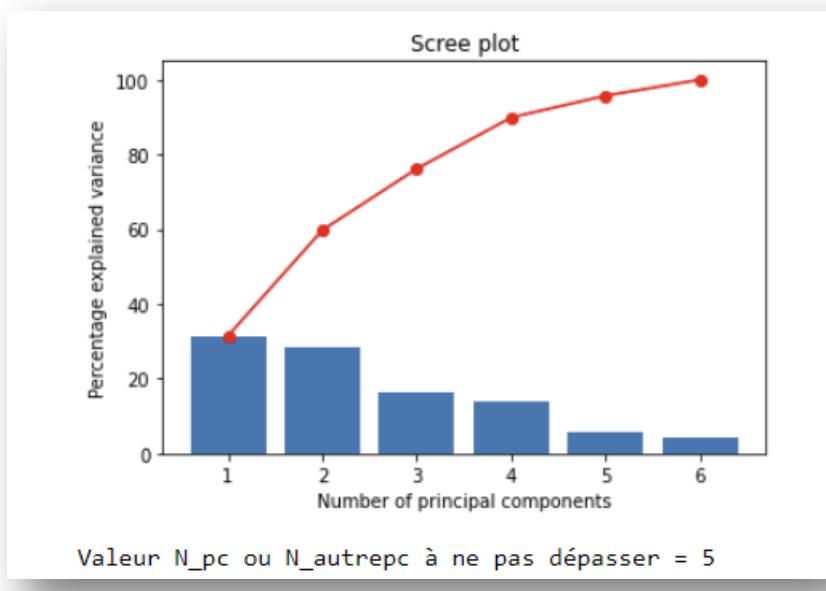
# Exploration des données\_Analyse Bi-Variée

- ▶ Boxplot avec une colonne catégorielle.



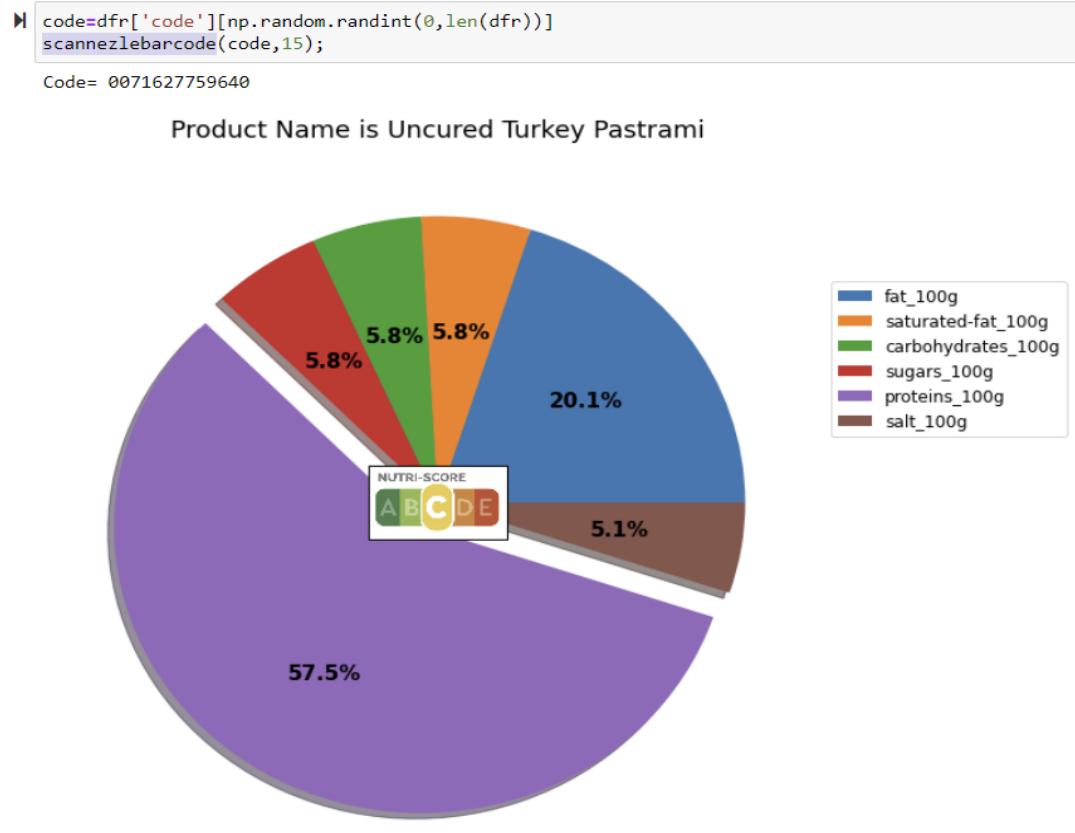
# Exploration des données\_Analyse Multi-Variée

- ▶ Présenter plus de 2 variables sur un plan 2d.
- ▶ ACP (Analyse en composante principale).
- ▶ La fonction analyse\_acp.



# Exemples d'application pertinentes

- ▶ Expliquer l'idée.
- ▶ La fonction scannezlebarcode.



# Conclusion

- ▶ Ce que j'ai appris.
- ▶ Mon sentiment après le 3<sup>ème</sup> projet.
- ▶ Gagner de vitesse et être mieux à l'aise permet de respecter les délais d'OpenClassroom.