Victor Benard

ML Engineer

Project 2

Cleaning and exploratory analysis of public nutrition data

Menu

- Data cleaning
  - Features selection
  - Application idea
  - Outliers
  - Missing data

- 2. Exploratory analysis
  - Univariate
  - Multivariate

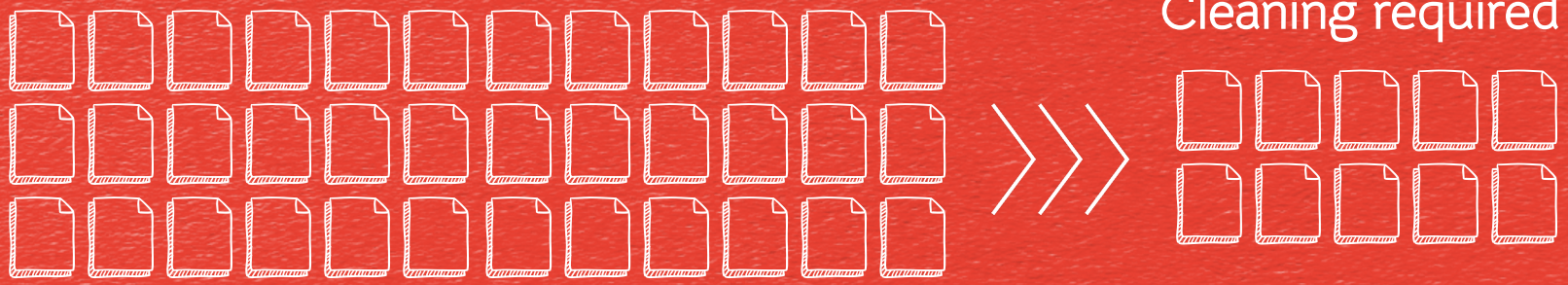
1.
Data cleaning

With pandas

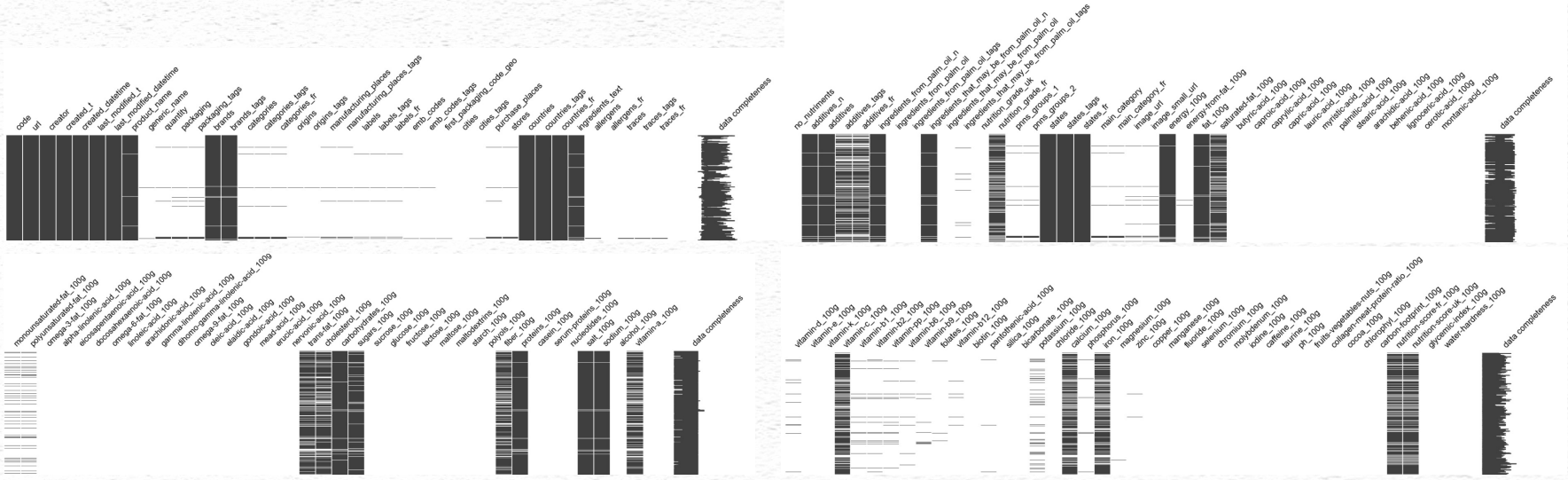
Dataframe size

df.shape

(320772, 162)

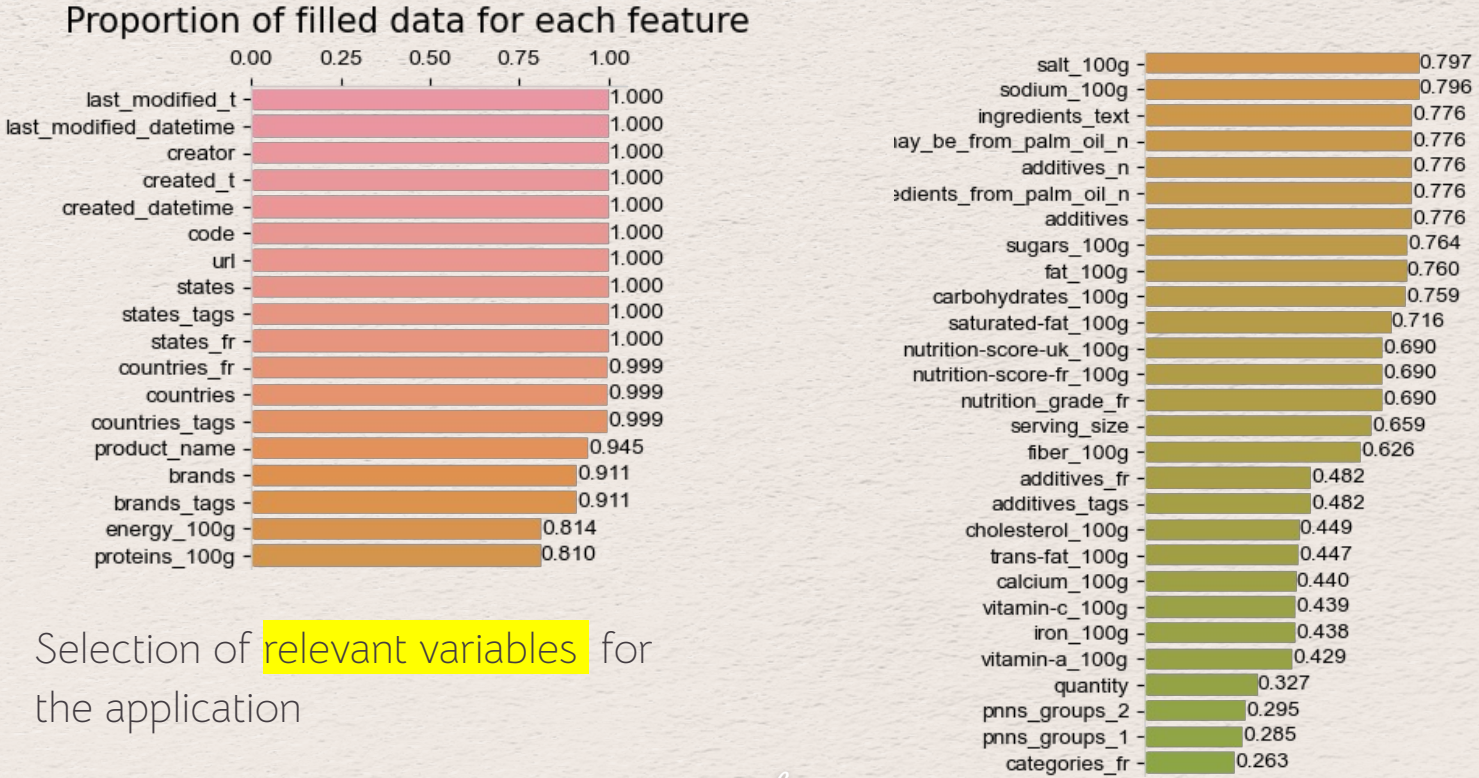


Missing data overview



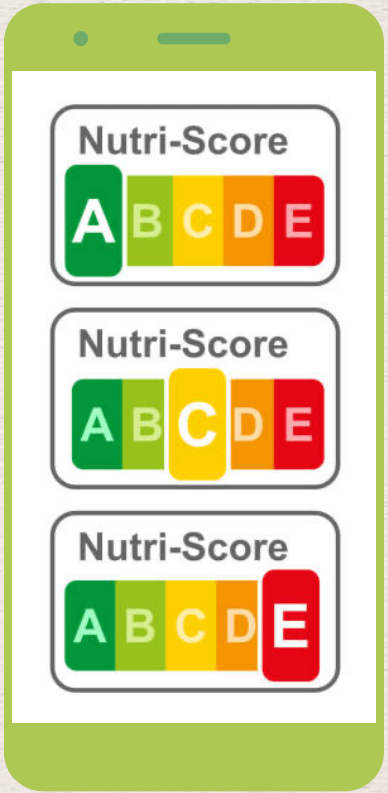
- 1 column out of 2 is empty or almost empty
- Vast majority of columns are incomplete

Missing data sorted by filling percengage



- Selection of relevant variables for the application

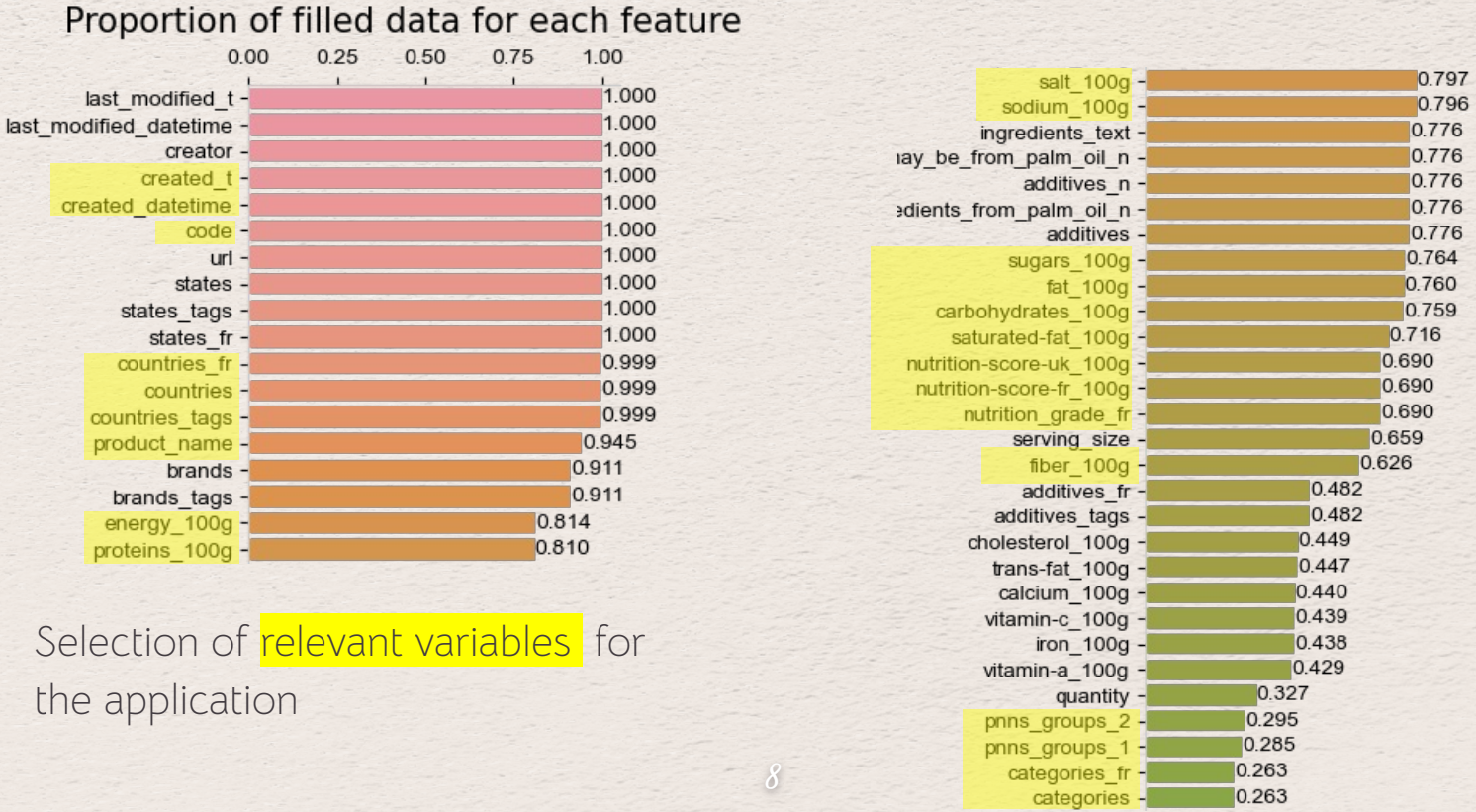
Application idea



Determine meals with an estimation of the best possible nutrition score possible.

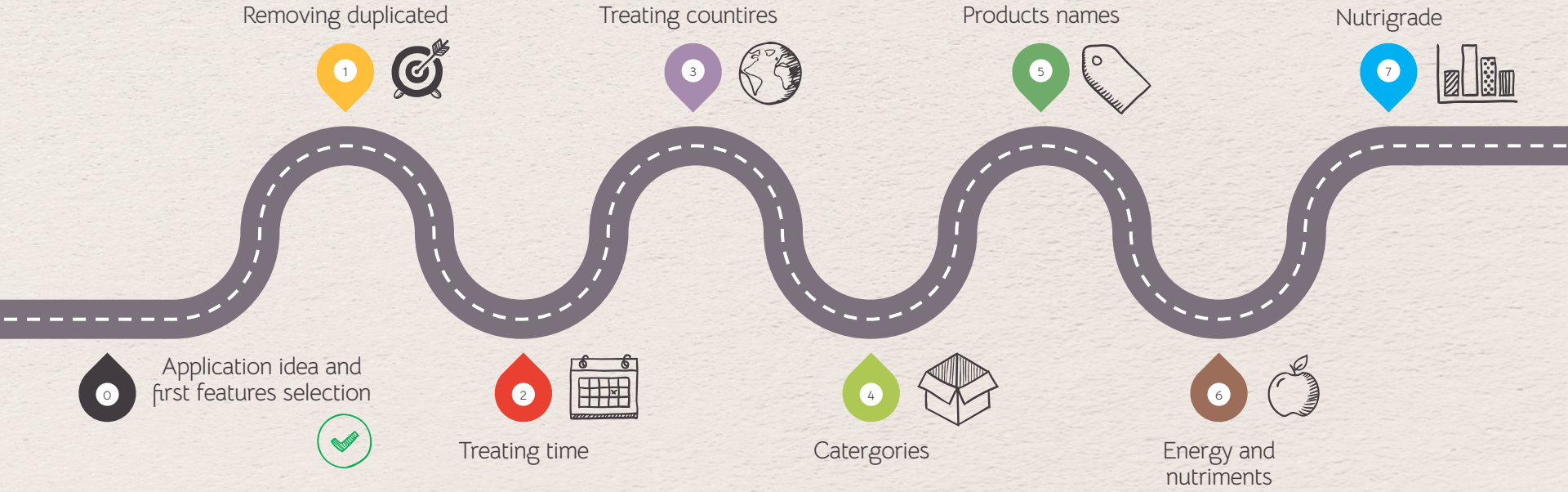
Application for France only.

Missing data sorted by filling percentage



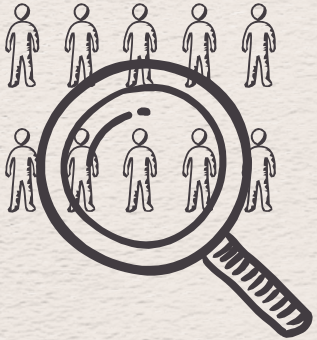
- Selection of relevant variables for the application

Cleaning roadmap



Removing duplicates

- Detection of duplicates through ‘code’ feature
- Counting number of NaN per row
- Sorting rows by missing values
- Dropping rows that contain more NaN



→ Removal of 133 duplicates

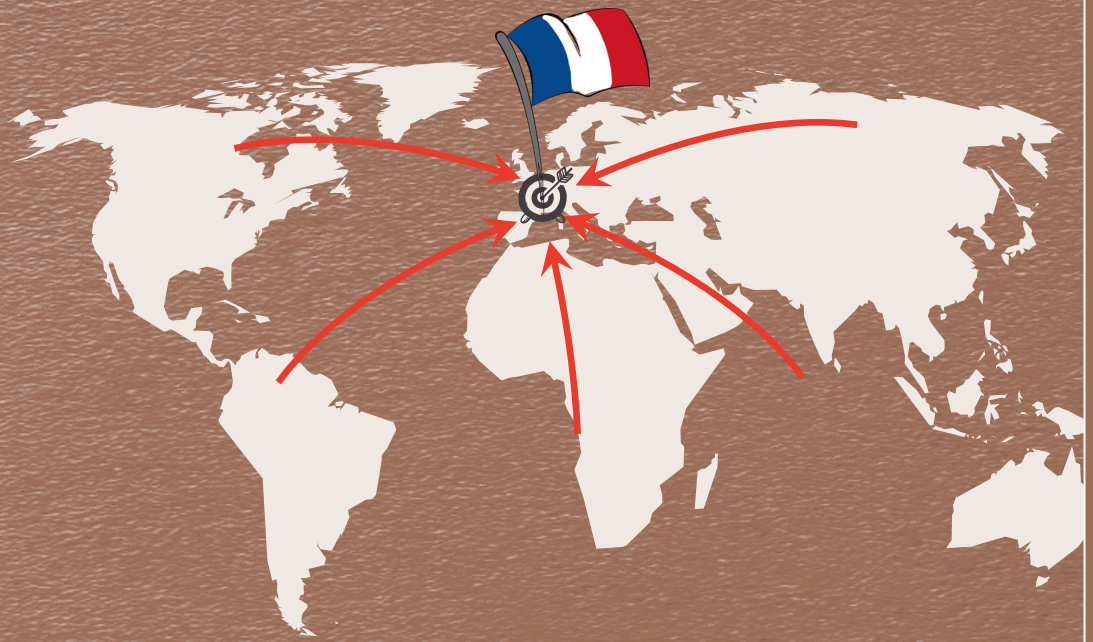
Treating of time

- Interesting only in creation date
- Two different features: created_t, created_datetime
  - Conversion in a common format

    pd.to_datetime(df[i], unit='s', errors='coerce')
  - Comparison
  - Deletion of the least well filled format

→ Conservation of 'created_datetime'

Treatment of countries



- Objective: France only
- 3 variables:
  - countries_fr
  - countries
  - countries_tags → retenu
- Formatting of tags
  - value_counts = 717
  - If several countries per row, keep only the first one
- Deletion of rows that do not concern France

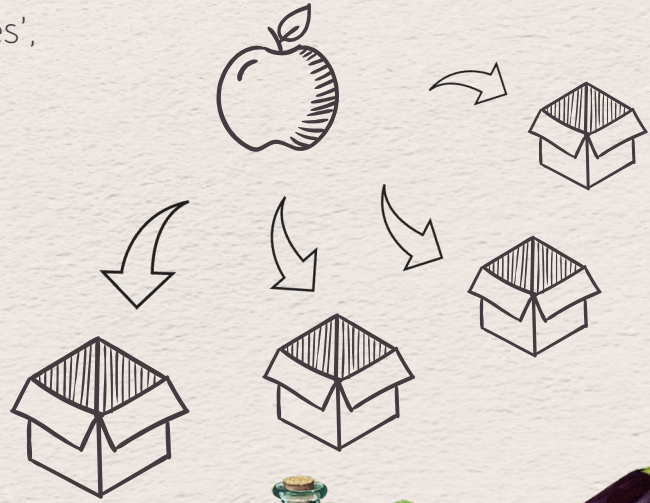
Categories

- Comparison of categories variables:

  'pnns_groups_1', 'pnns_groups_2', 'categories_fr', 'categories', 'categories_Tags', 'main_category', 'main_category_fr'

- Selection of 'pnns_groups_1'
- Replacing redundant values through a dictionary

  ex: {'sugary-snacks':'Sugary snacks'}



Products names

- Objective : have a product name per row
- Detection and deletion of missing date, using the feature ‘product_name’
- 4 produit out of 5 makes a unique appearance

Energy and nutriments

Energy

- .describe() shows values over 3700kJ, that can't be
- Replacement of outliers by the median value

Energy and nutriments

Nutriments

- Concerns:

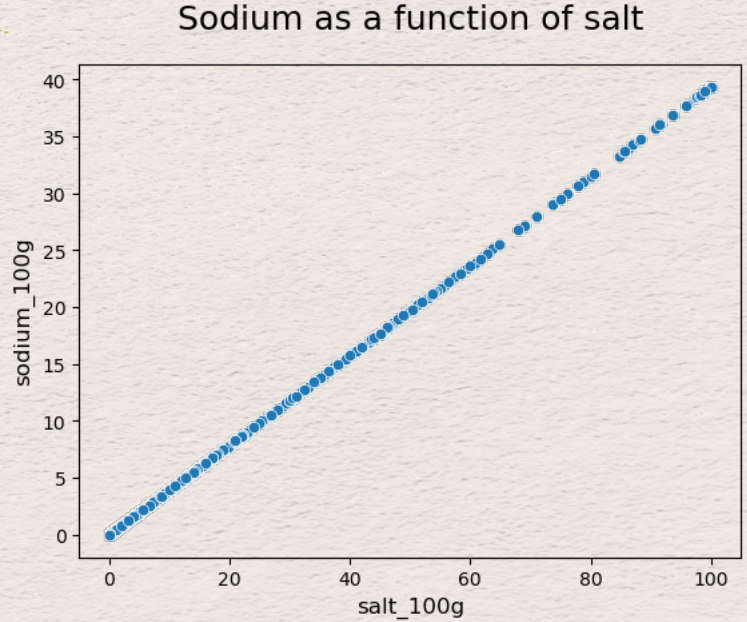
  'proteins_100g','salt_100g','sugars_100g','fat_100g','carbohydrates_100g','saturated-fat_100g','fiber_100g','fruits-vegetables-nuts_100g','sodium_100g'

- Replacing negative values by 0
- Replacing values > 100g by NaN

Energy and nutriments

Salt or Sodium?

- Comparison between salt and sodium
- Deletion of sodium feature



Energy and nutriments

Sum of nutriments < 100g

- Sum of columns

'carbohydrates_100g','proteins_100g','fat_100g','fiber_100g'

- Deletion of rows where the total exceeds 100g

Nutriments with NaN per row < 3

- Concerns

'energy_100g','sugars_100g','saturated-fat_100g','fruits-vegetables-nuts_100g', 'fiber_100g', 'proteins_100g','salt_100g','fat_100g','carbohydrates_100g'

- Counting number of filled values per row
- Suppression of rows with 3 NaN or more

Energy and nutriments

Remplissage des nutriments

- Pour ‘fiber_100g’: NaN remplacés par 0
- Pour les autres nutriments:
  - Groupement par catégorie
  - Remplissage par la médiane associée à chaque catégorie

Nutrigrade

Determination by KNN imputer

- One-hot encoding of categories of ‘pnns_groups_1’ and ‘nutrition_grade_fr’
- Training on following variables:

'energy_100g','sugars_100g','saturated-fat_100g','salt_100g',
'fruits-vegetables-nuts_100g', 'fiber_100g', 'proteins_100g',
'carbohydrates_100g','fat_100g' + one-hot encoding

Cleaned dataframe

61K rows

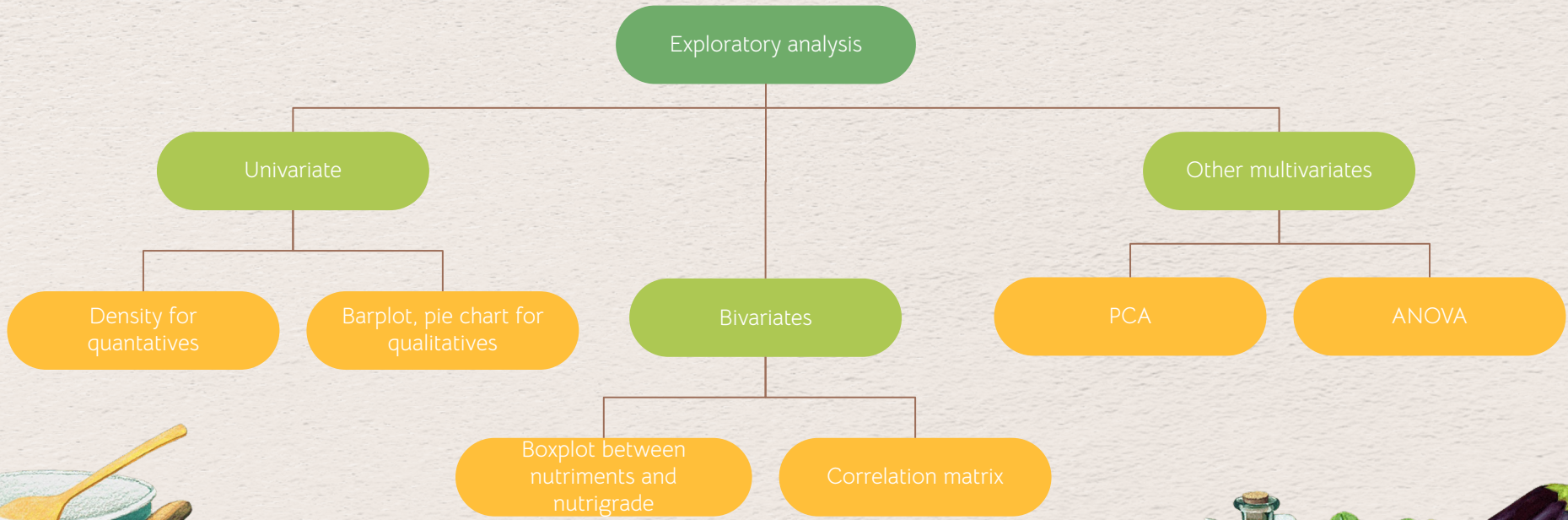
15 columns

2.

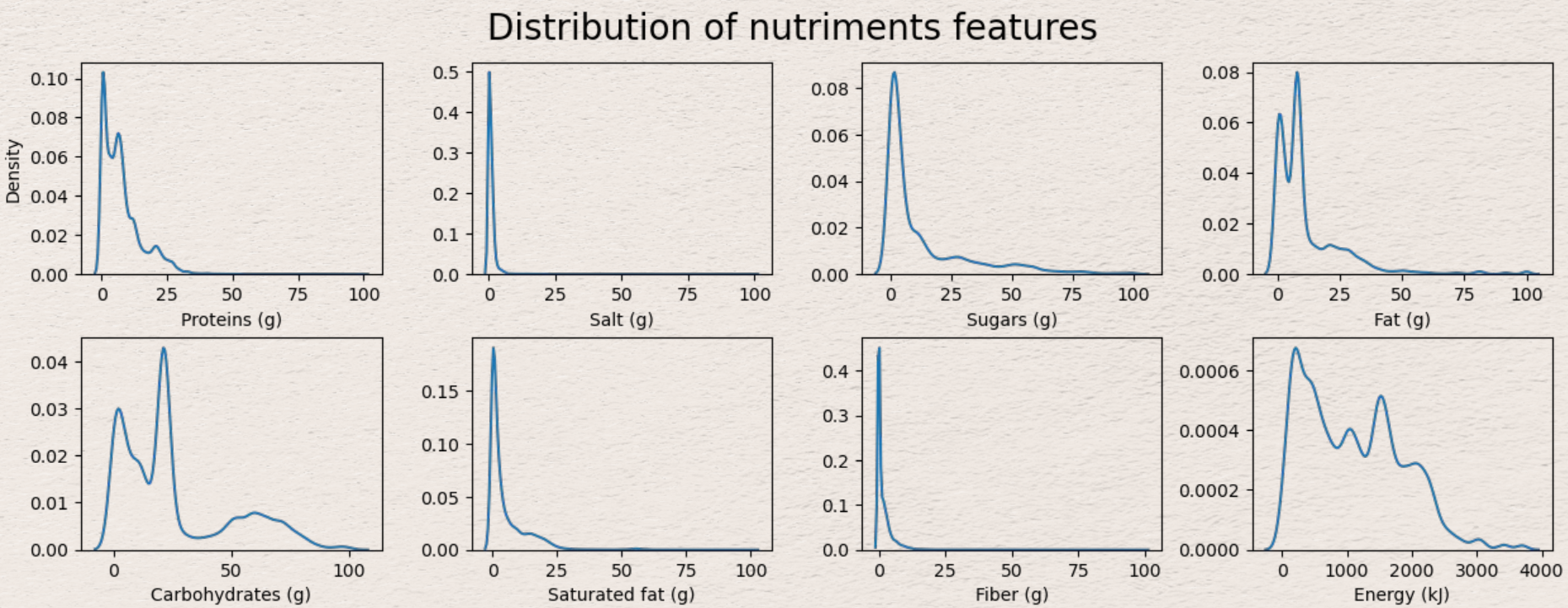
Exploratory analysis

With pandas, matplotlib, and seaborn

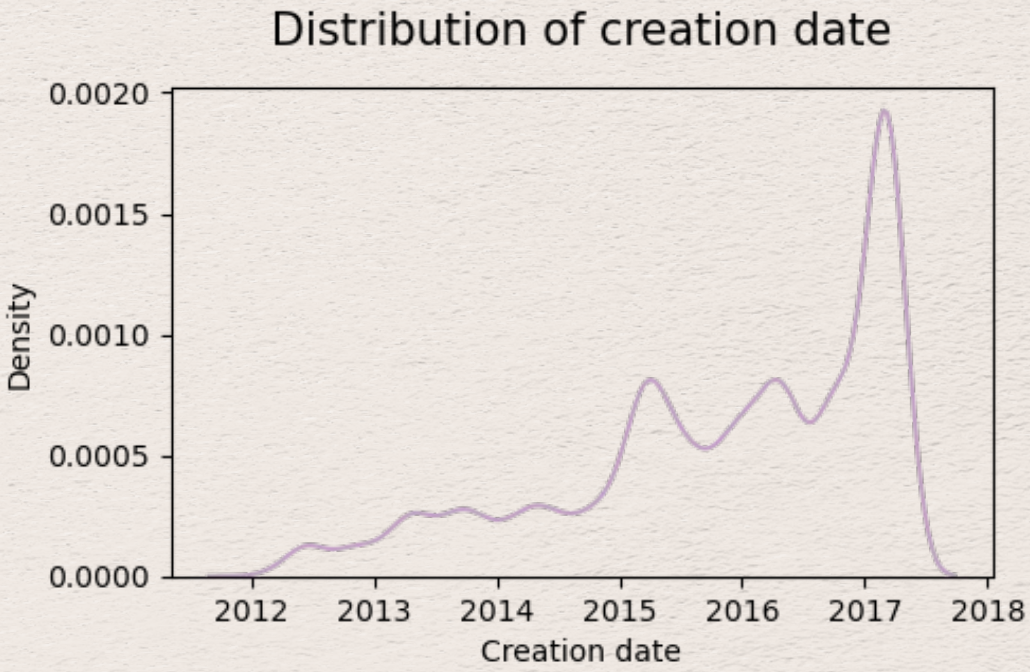
Analysis types



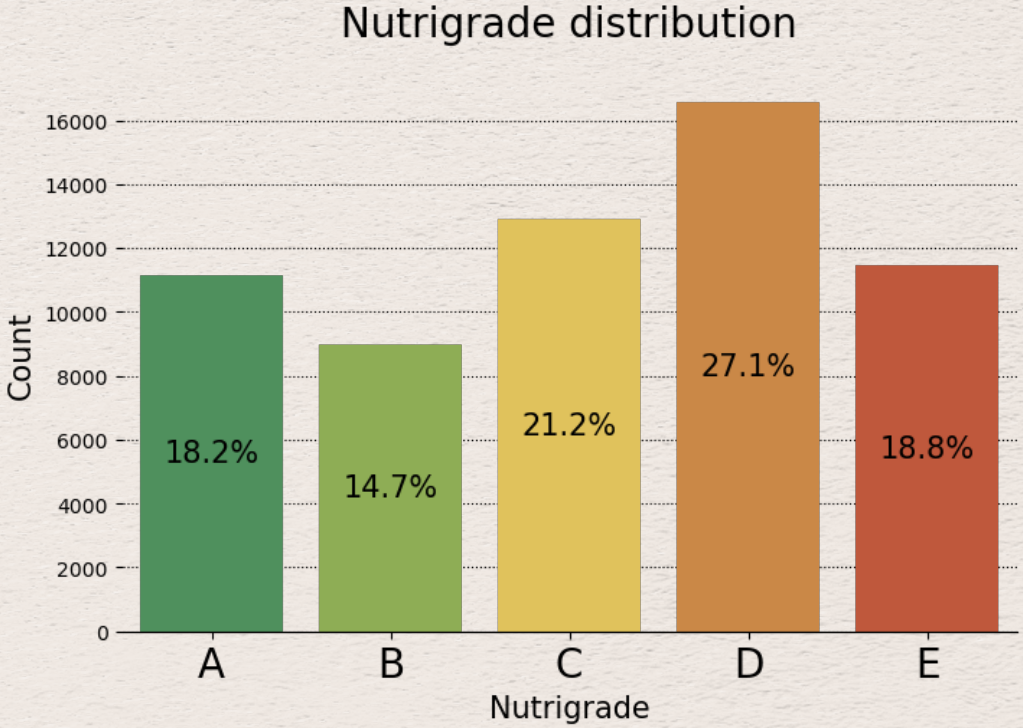
Univariate analysis



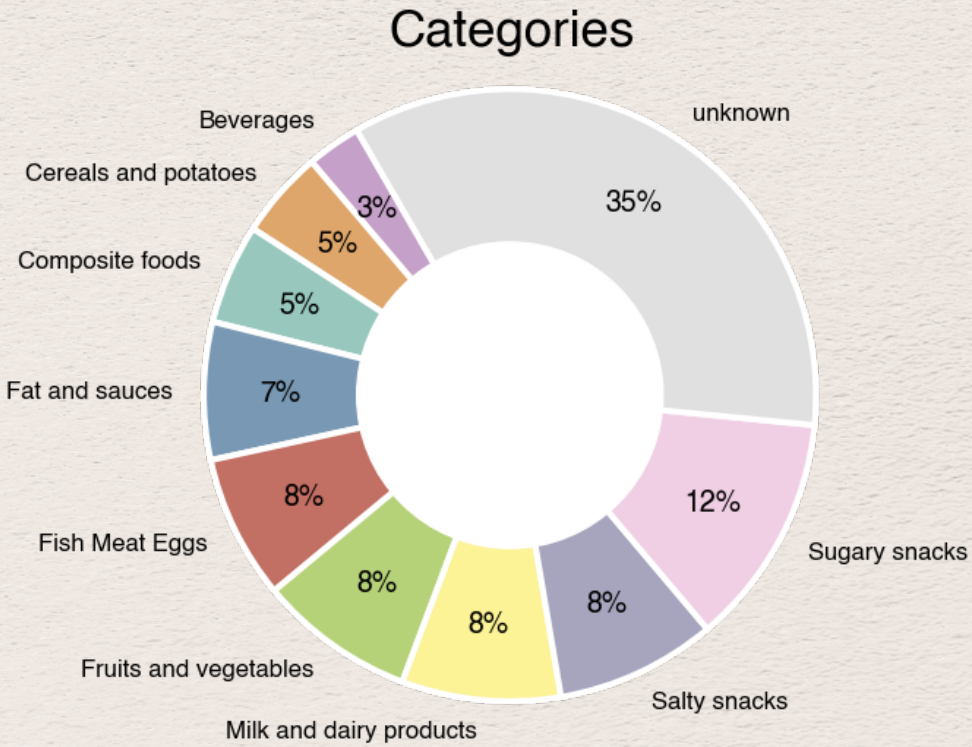
Univariate analysis



Univariate analysis

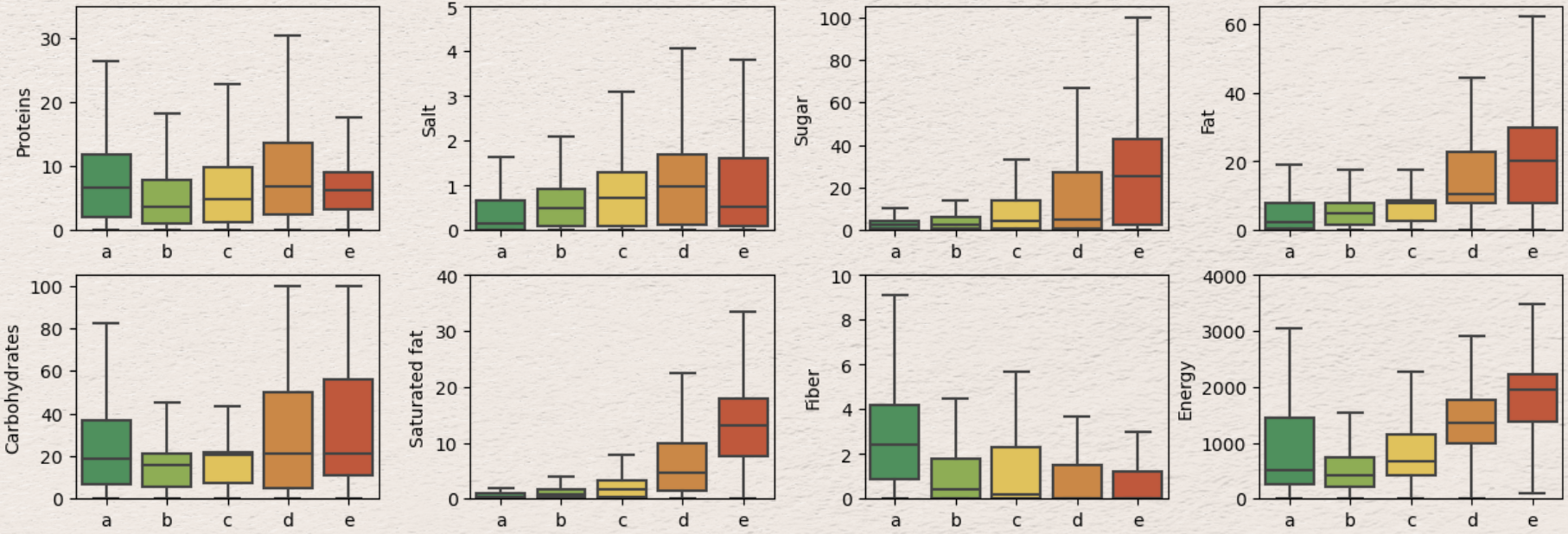


Univariate analysis



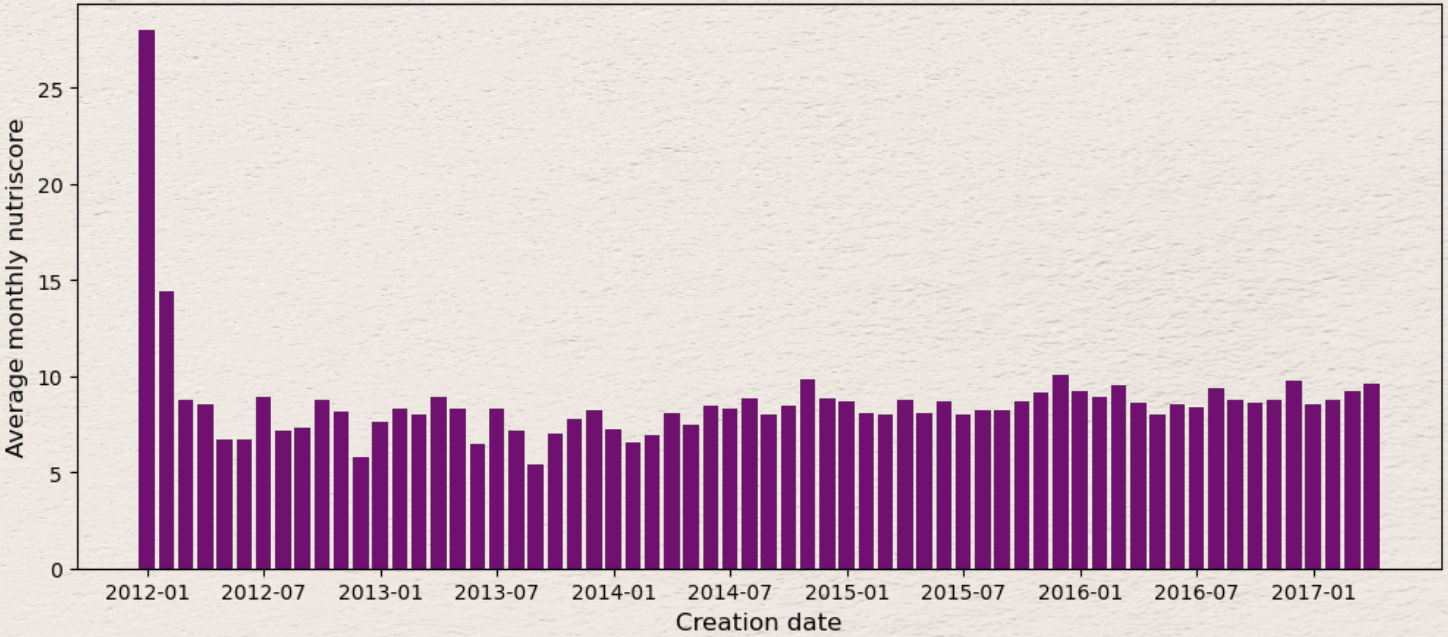
Bivariate analysis

Distribution of nutriments in relation to nutrigrade



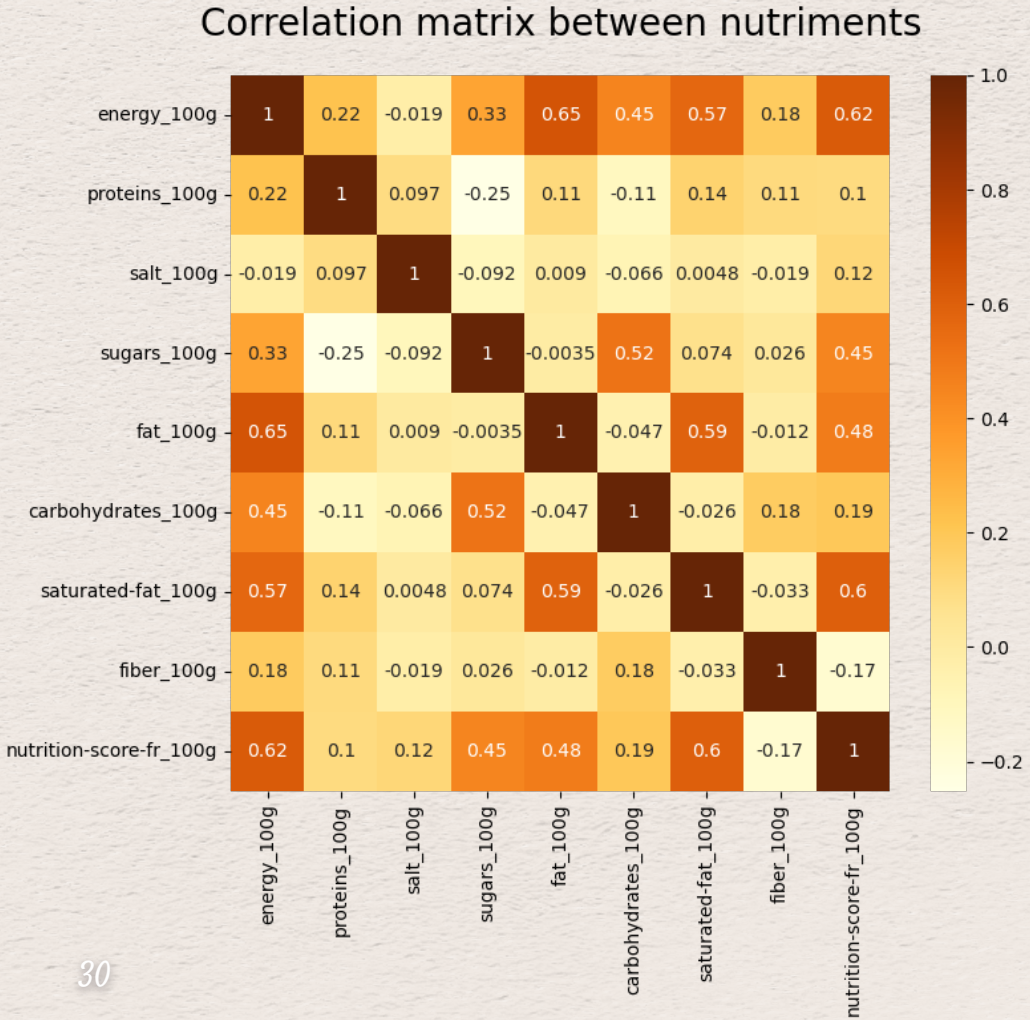
Bivariate analysis

Nutriscore evolution along time

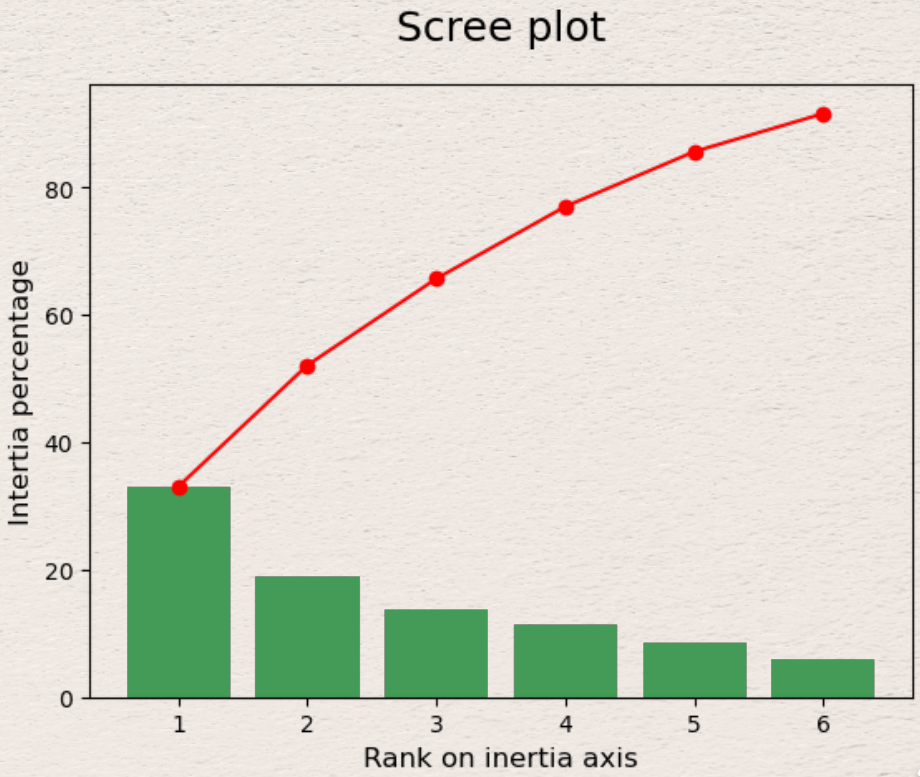


Bivariate analysis

- Observed correlations:
  - Energy: fat, proteins, carbs
  - Carbs: sugar, fiber
  - Fat: saturated fat
- Anti-correlations:
  - Proteins: sugar



Multivariate analysis



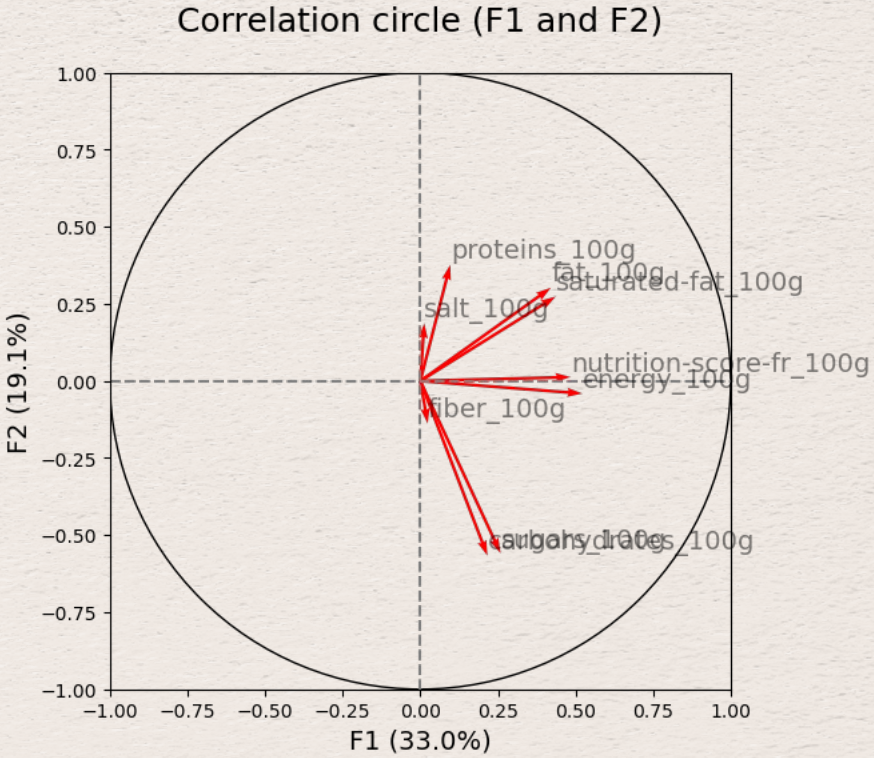
Principal component analysis

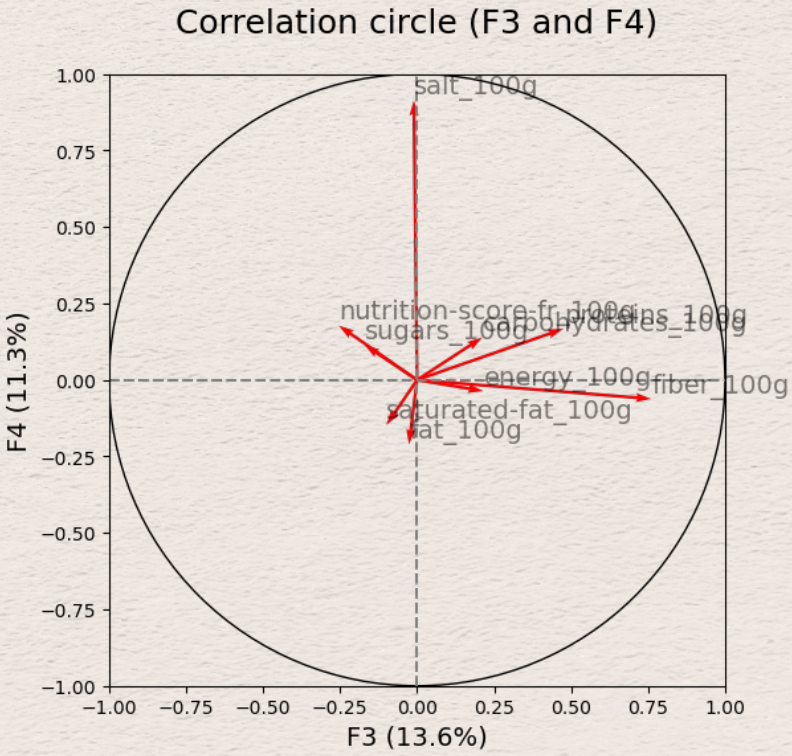
- Concerns

'energy_100g','proteins_100g','salt_100g','sugars_100g','fat_100g', 'carbohydrates_100g','saturated-fat_100g','fiber_100g','nutrition-score-fr_100g'

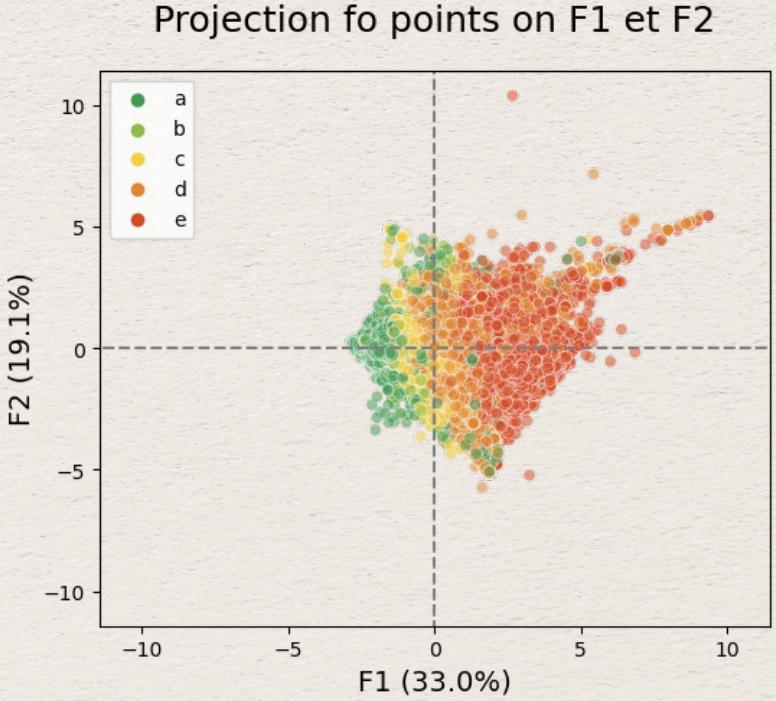
- Normalization
- Principal components calculation
- Scree plot
- More than 50% of inertia is on the 2 first axis

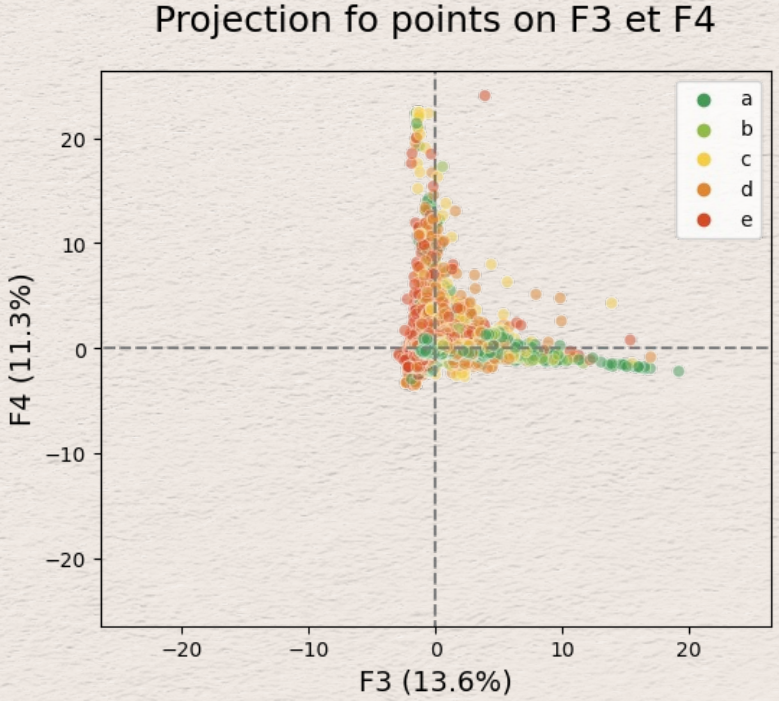
Multivariate analysis





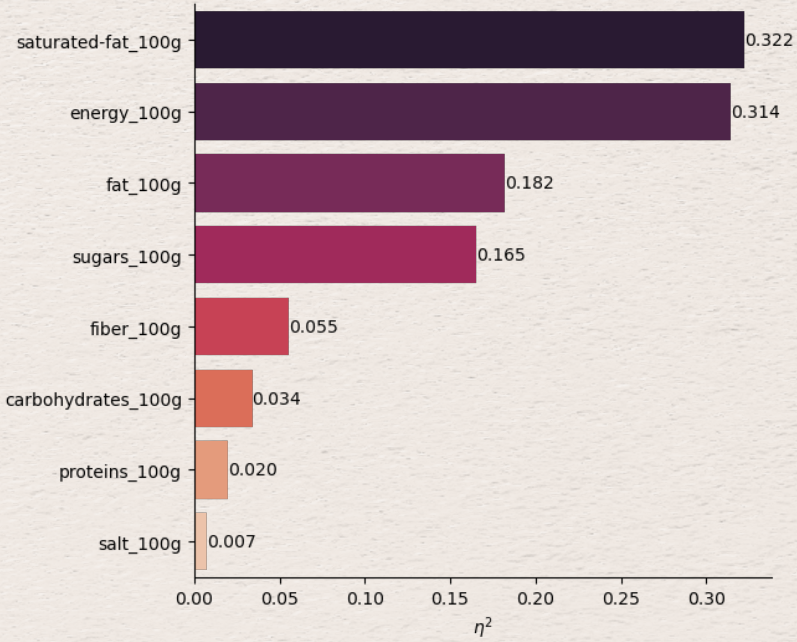
Multivariate analysis



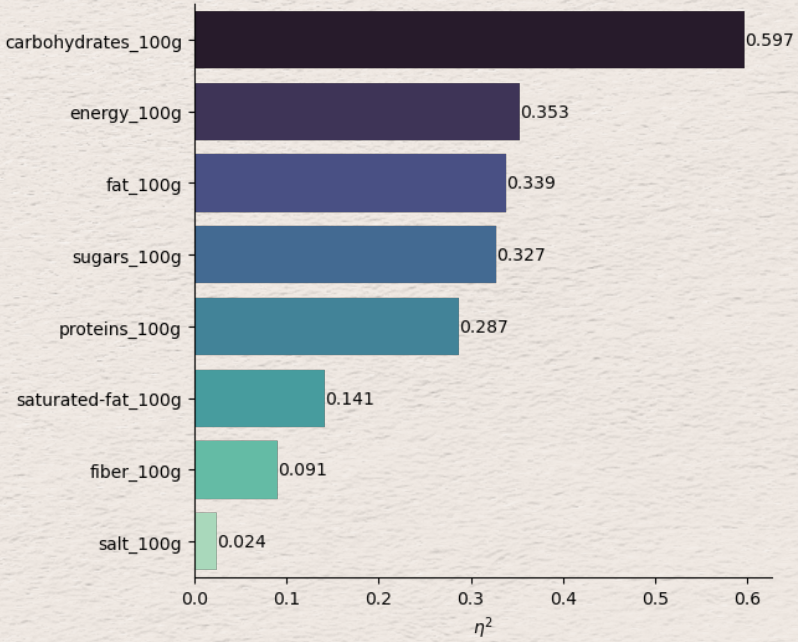


Multivariate analysis

ANOVA between nutriments and nutrigrade



ANOVA between nutriments and catégories



Thank you !

Do you have any questions?