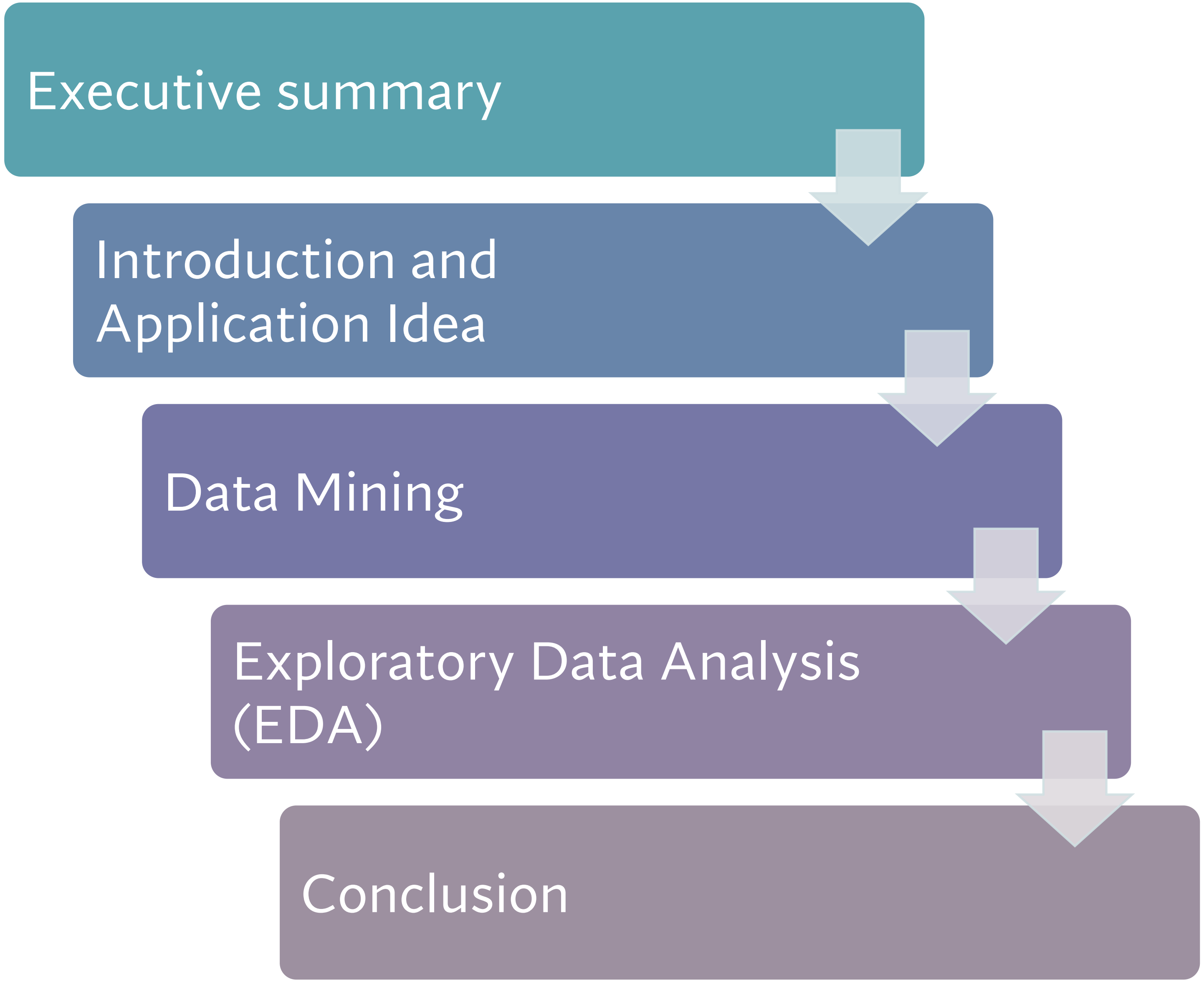
Health Application Project: Alimentation Médicale Préventive

Adonija ZIO

October 2022

Outline



Executive summary

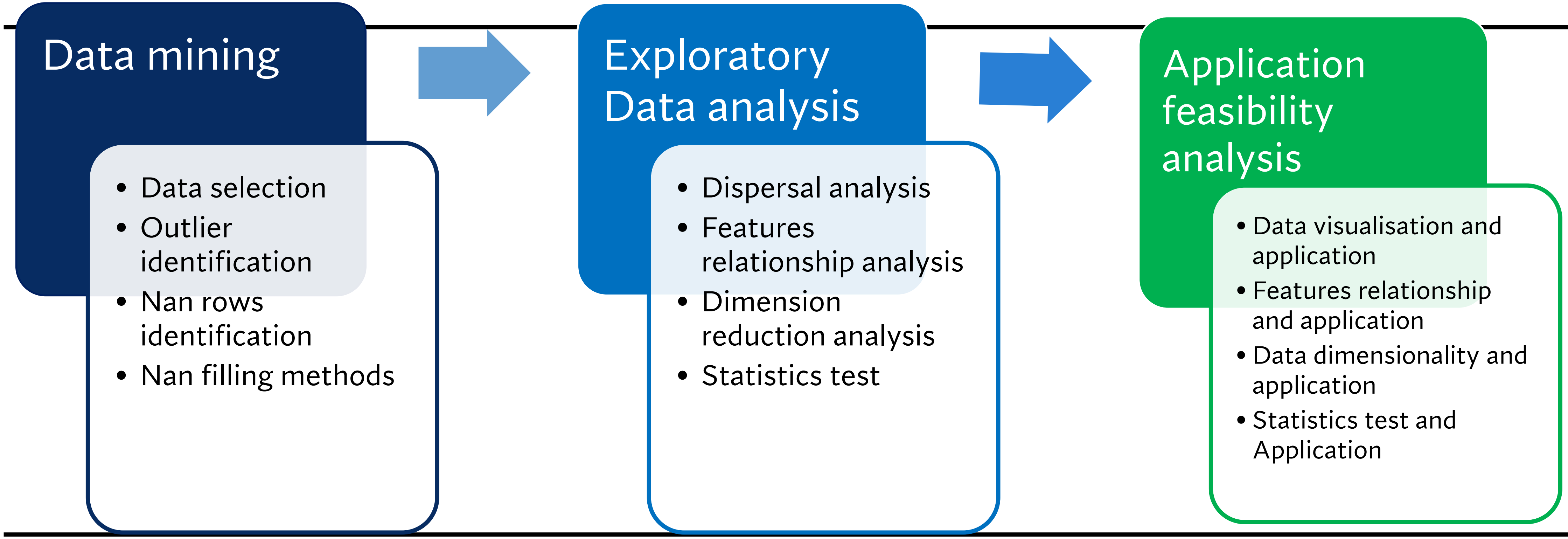
Introduction and Application Idea

Data Mining

Exploratory Data Analysis (EDA)

Conclusion

Executive summary



Data mining

- Data selection
- Outlier identification
- Nan rows identification
- Nan filling methods

Exploratory Data analysis

- Dispersal analysis
- Features relationship analysis
- Dimension reduction analysis
- Statistics test

Application feasibility analysis

- Data visualisation and application
- Features relationship and application
- Data dimensionality and application
- Statistics test and Application

Introduction

Steady increase in diabetes cases

463 million adults (20-79 years) with diabetes worldwide in 2019

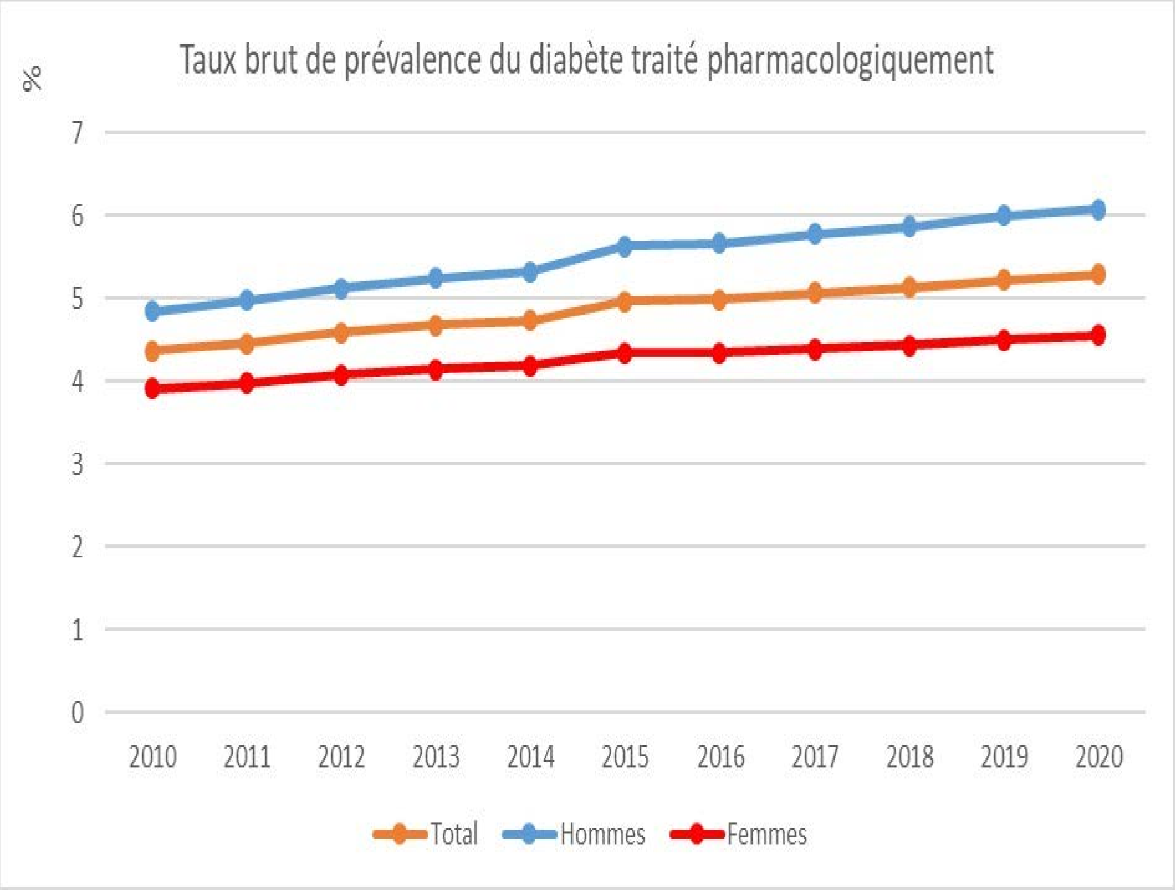
This is expected to rise to **700 million by 2045**

Over **4 million people** identified in France in 2019

3,5 millions under treatment in France in 2020

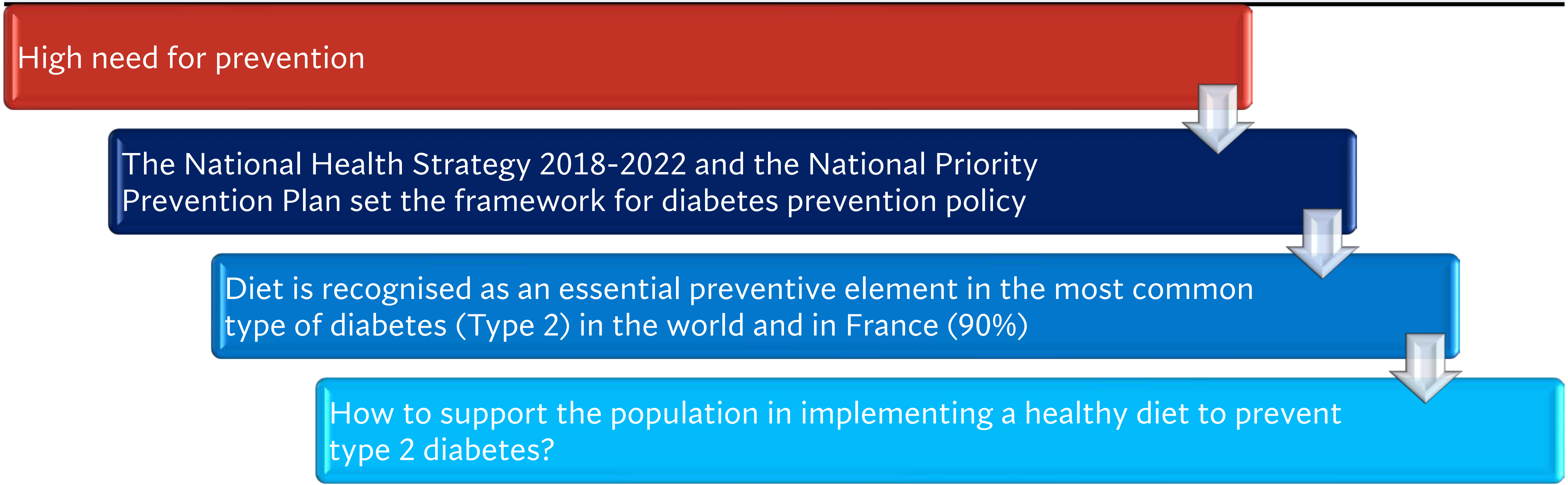
10% des diabétiques sont de type 1

These treatments are very high cost for the French health system (**19 billion** per year according to CNAM)



Source: Santé Publique France

Introduction



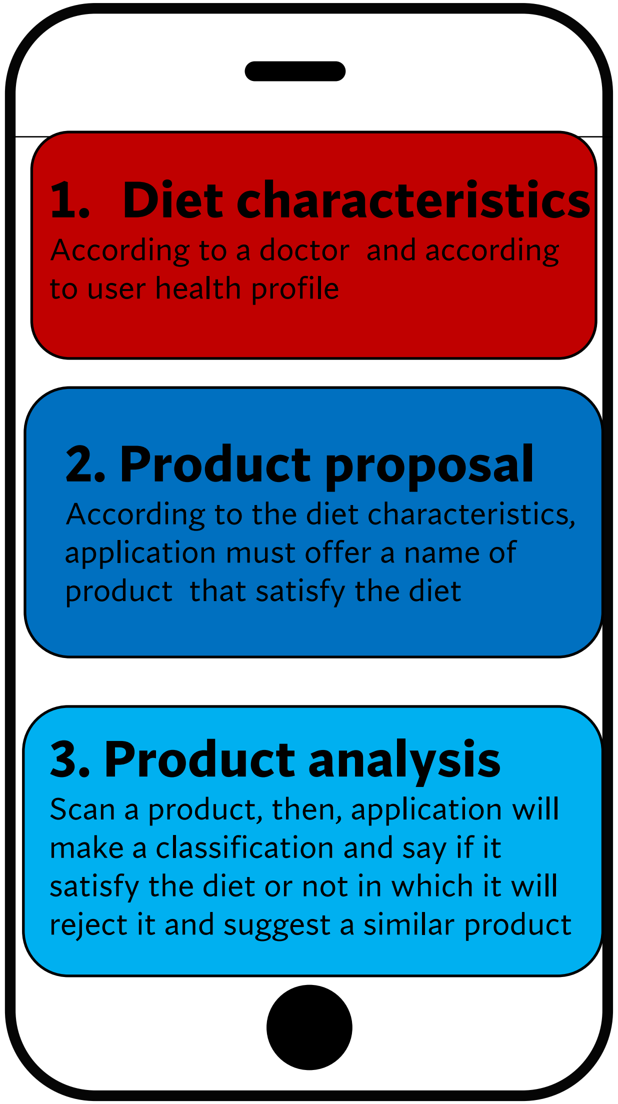
High need for prevention

The National Health Strategy 2018-2022 and the National Priority Prevention Plan set the framework for diabetes prevention policy

Diet is recognised as an essential preventive element in the most common type of diabetes (Type 2) in the world and in France (90%)

How to support the population in implementing a healthy diet to prevent type 2 diabetes?

Application Idea



1. Diet characteristics

According to a doctor and according to user health profile

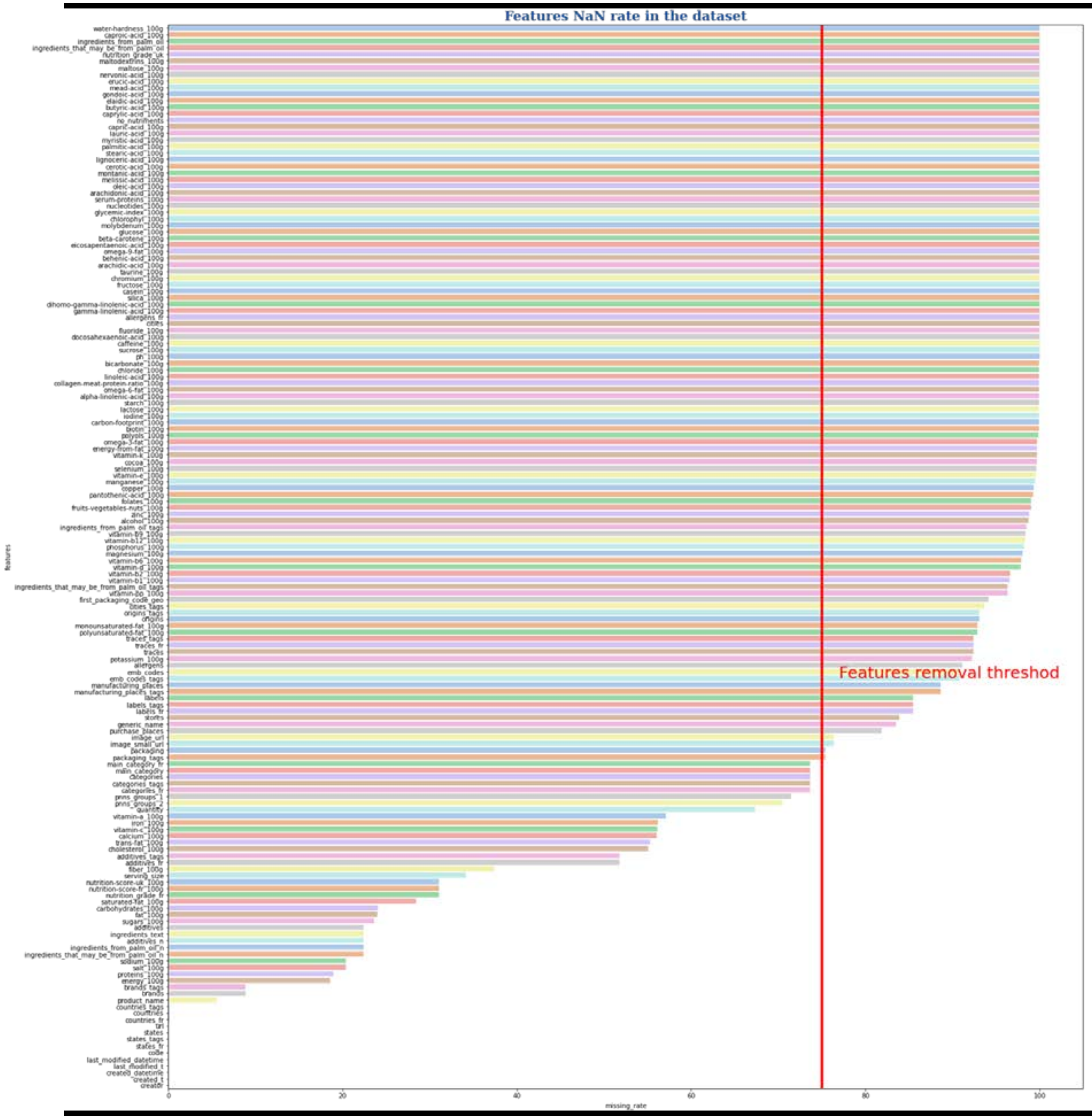
2. Product proposal

According to the diet characteristics, application must offer a name of product that satisfy the diet

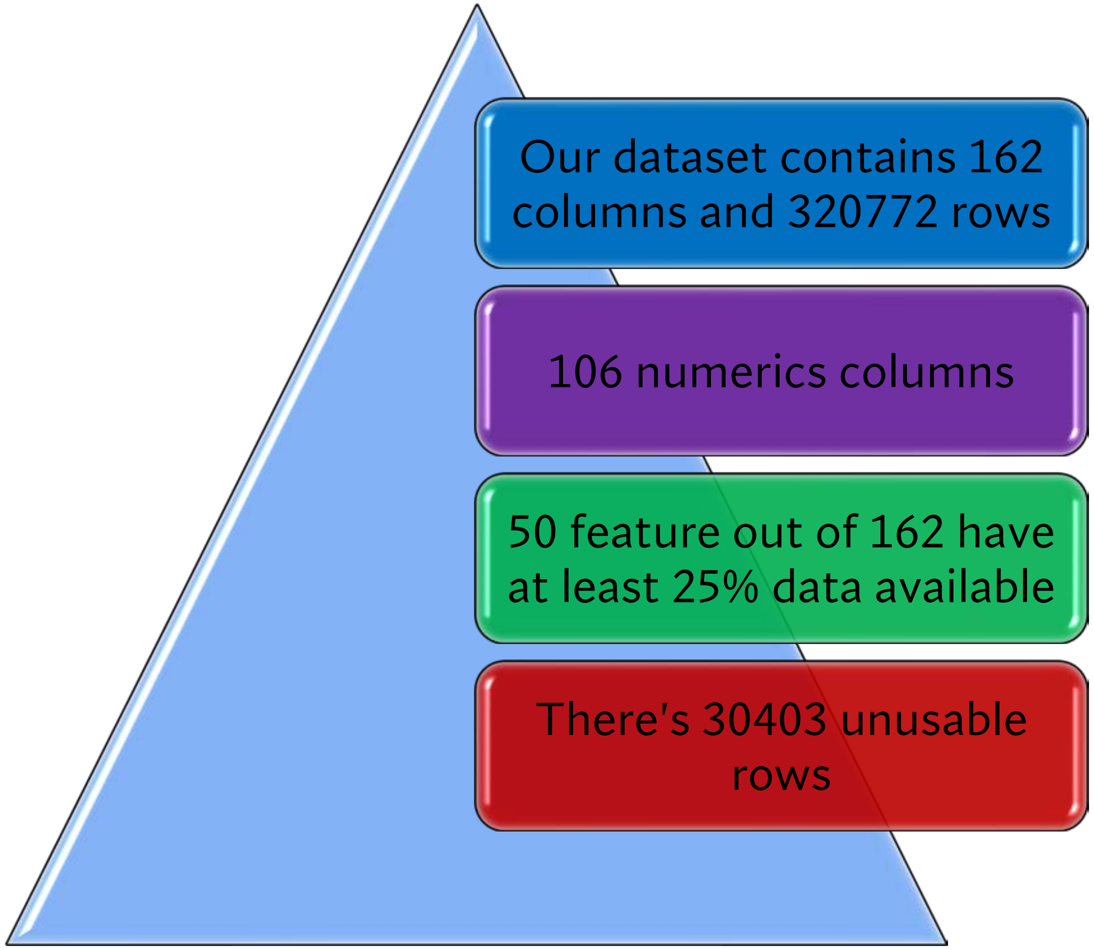
3. Product analysis

Scan a product, then, application will make a classification and say if it satisfy the diet or not in which it will reject it and suggest a similar product

Data Mining



Technical selection



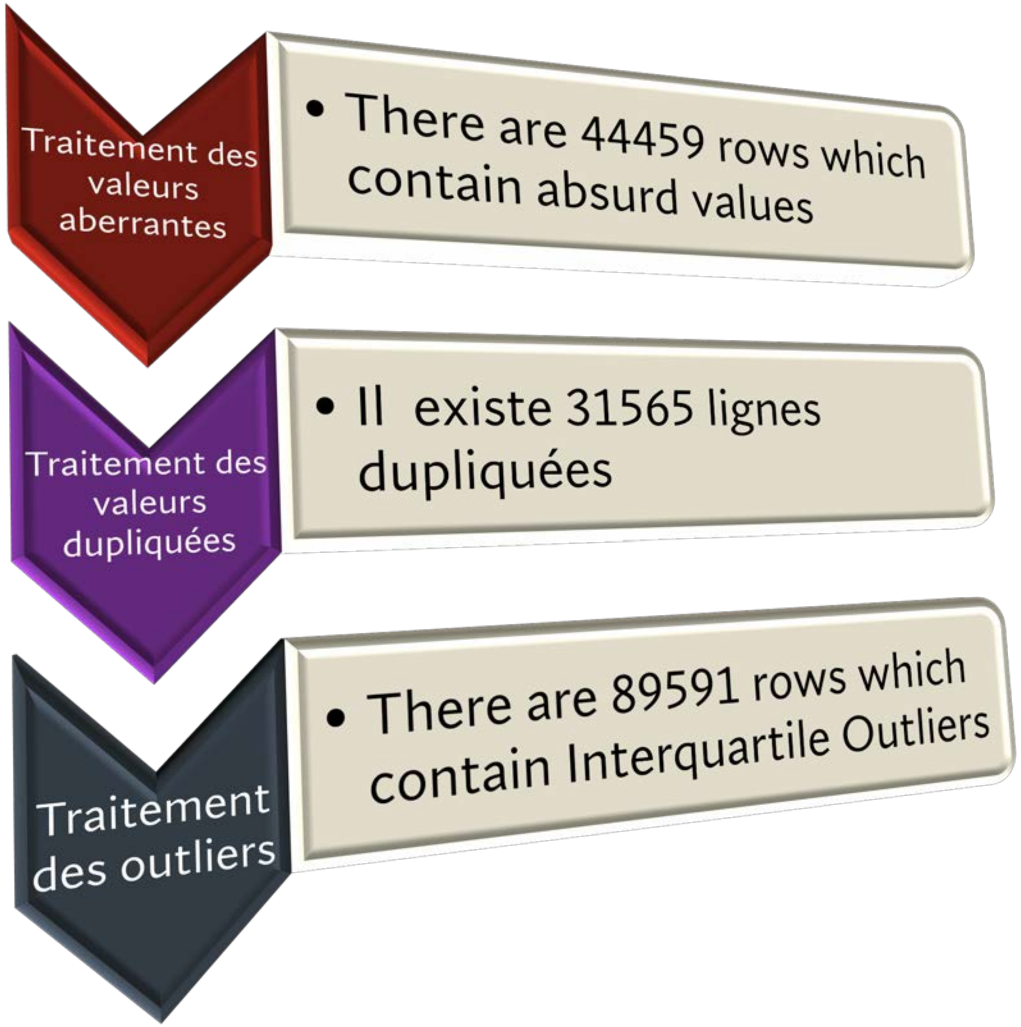
Our dataset contains 162 columns and 320772 rows

106 numerics columns

50 feature out of 162 have at least 25% data available

There's 30403 unusable rows

Value treatment



Traitement des valeurs aberrantes

- There are 44459 rows which contain absurd values

Traitement des valeurs dupliquées

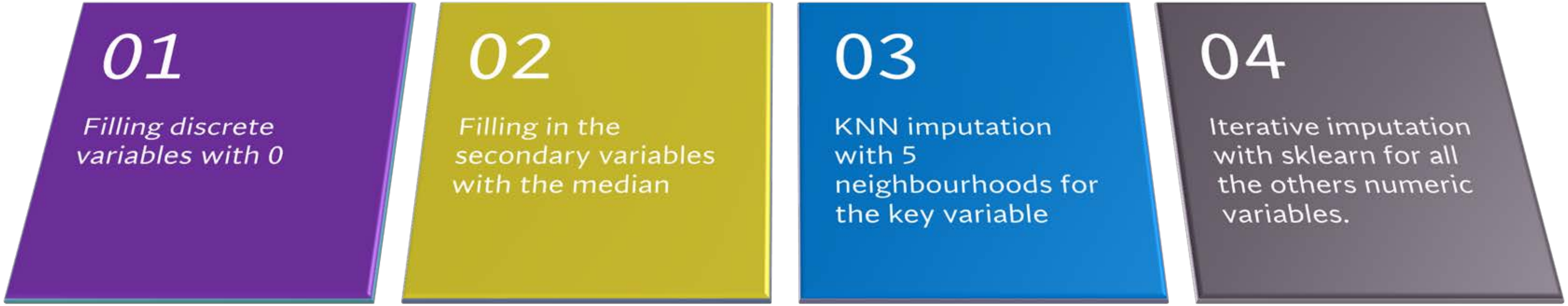
- Il existe 31565 lignes dupliquées

Traitement des outliers

- There are 89591 rows which contain Interquartile Outliers

Inter Quartile Range for energy_100g is 1284.0
Inter Quartile Range for proteins_100g is 9.41
Inter Quartile Range for salt_100g is 1.40602
Inter Quartile Range for sodium_100g is 0.480811023622047
Inter Quartile Range for additives_n is 3.0
Inter Quartile Range for sugars_100g is 27.41
Inter Quartile Range for fat_100g is 22.02
Inter Quartile Range for carbohydrates_100g is 50.33
Inter Quartile Range for saturated-fat_100g is 7.875
Inter Quartile Range for nutrition-score-fr_100g is 13.0
Inter Quartile Range for nutrition-score-uk_100g is 14.0
Inter Quartile Range for fiber_100g is 3.3
Inter Quartile Range for cholesterol_100g is 0.025
Inter Quartile Range for trans-fat_100g is 0.0
Inter Quartile Range for calcium_100g is 0.089
Inter Quartile Range for vitamin-c_100g is 0.005
Inter Quartile Range for iron_100g is 0.00203
Inter Quartile Range for vitamin-a_100g is 8.34e-05

Nan Filling



01

Filling discrete variables with 0

02

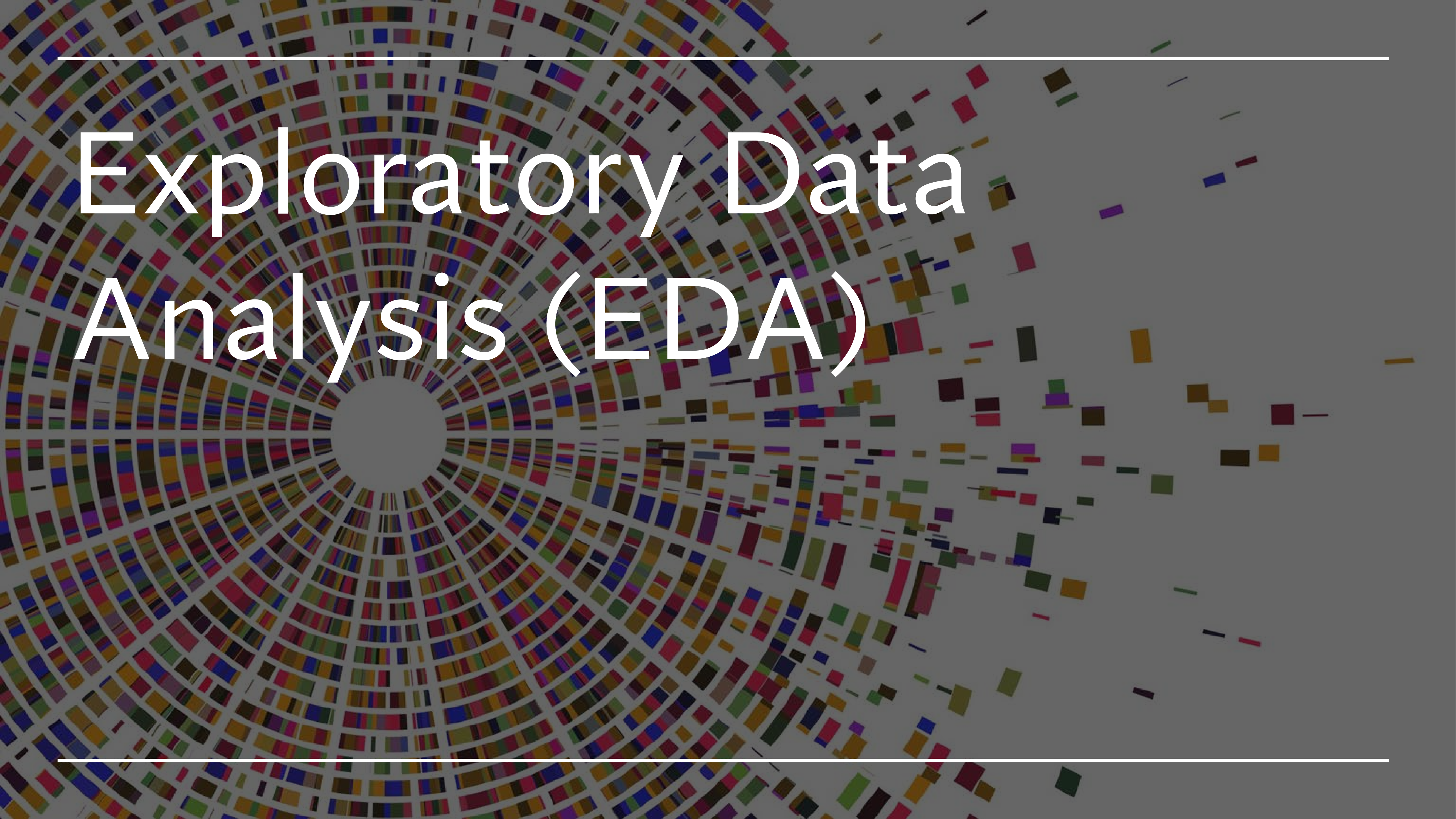
Filling in the secondary variables with the median

03

KNN imputation with 5 neighbourhoods for the key variable

04

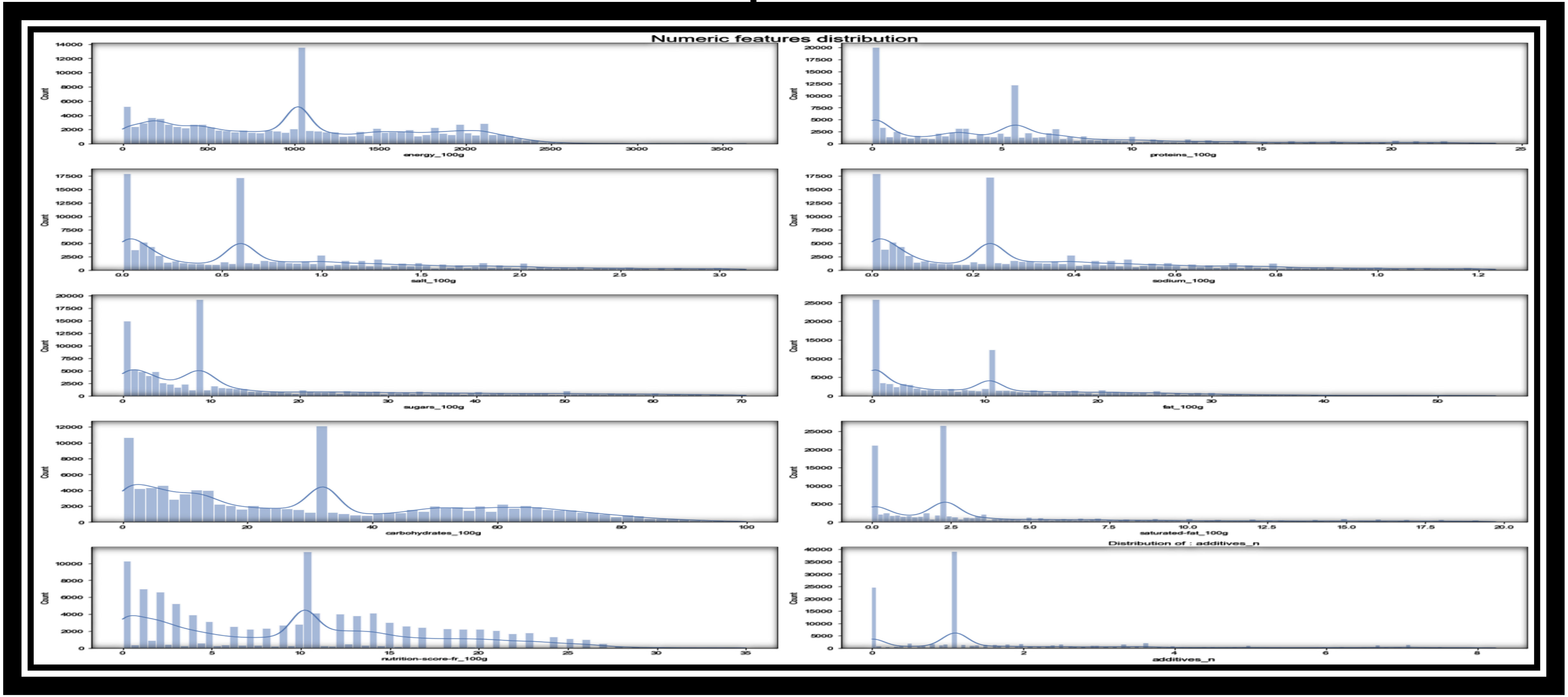
Iterative imputation with sklearn for all the others numeric variables.



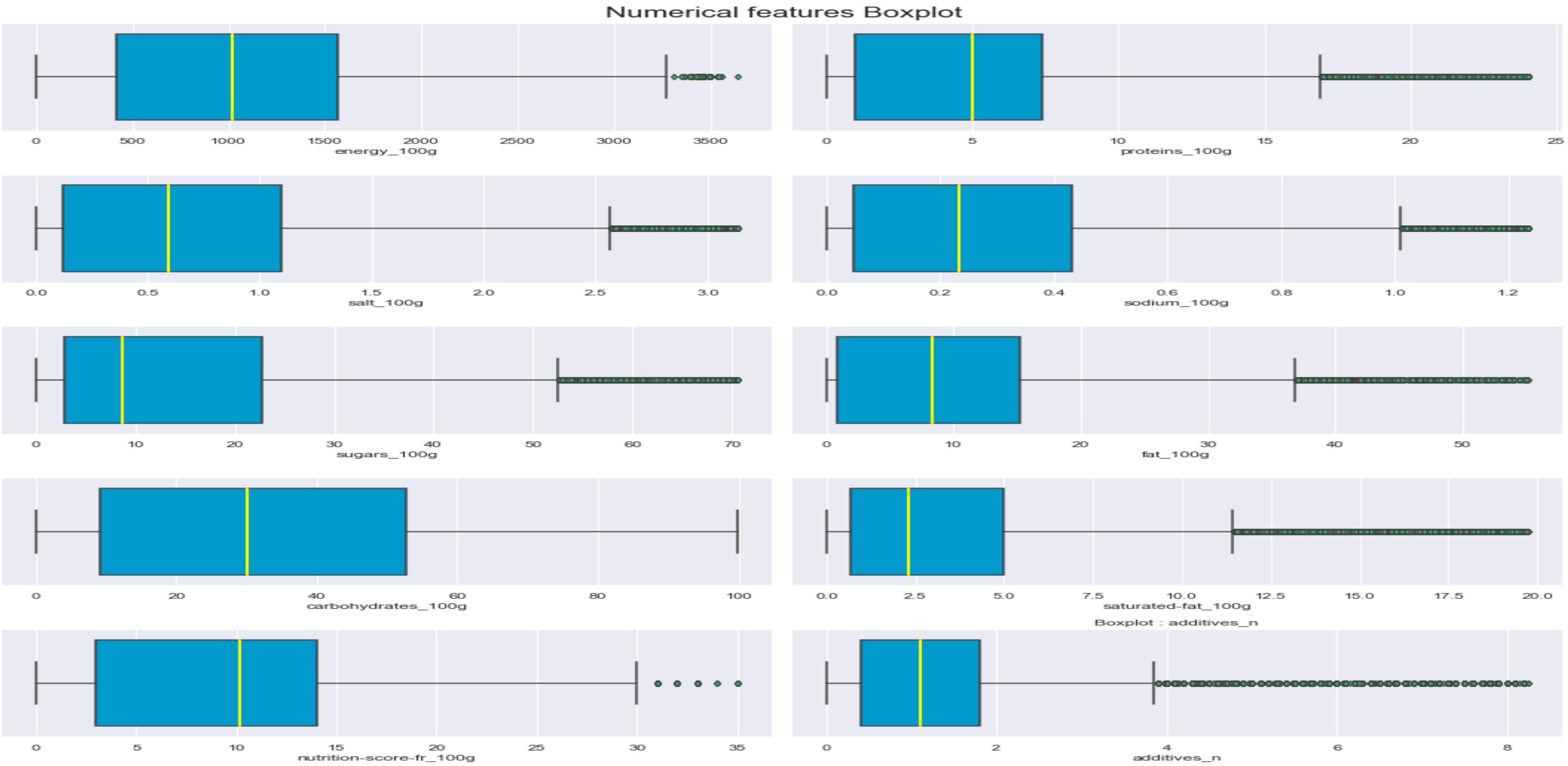
Exploratory Data Analysis (EDA)

Univariate analysis

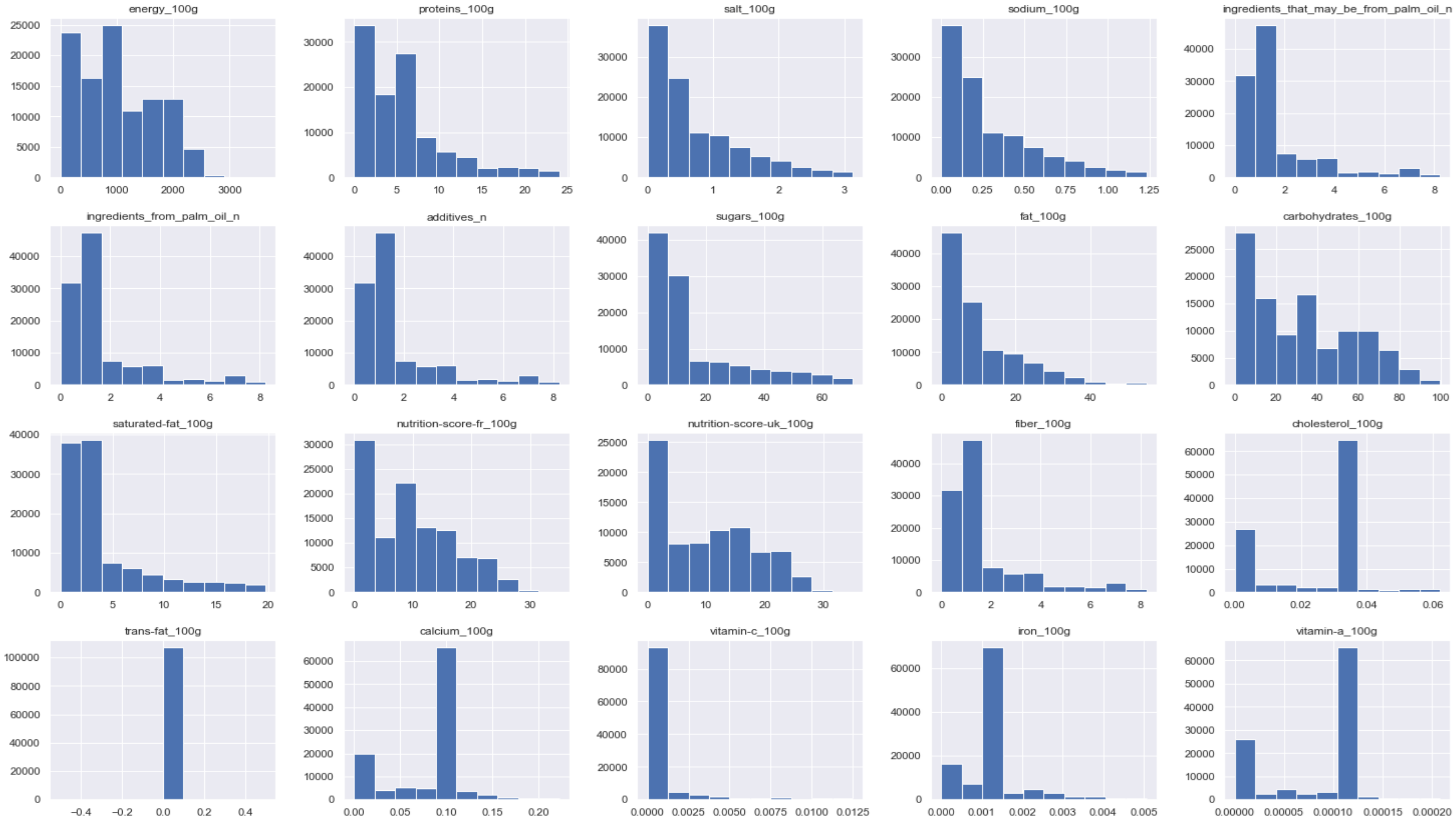
Displot


Numeric features distribution
Count
energy_100g
0
500
1000
1500
2000
2500
3000
3500
0
2000
4000
6000
8000
10000
12000
14000
proteins_100g
0
5
10
15
20
25
2500
5000
7500
10000
12500
15000
17500
20000
salt_100g
0.0
0.5
1.0
1.5
2.0
2.5
3.0
sodium_100g
0.0
0.2
0.4
0.6
0.8
1.0
1.2
sugars_100g
0
10
20
30
40
50
60
70
fat_100g
0
10
20
30
40
50
25000
carbohydrates_100g
0
20
40
60
80
100
saturated-fat_100g
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
nutrition-score-fr_100g
0
5
10
15
20
25
30
35
Distribution of : additives_n
additives_n
0
2
4
6
8
30000
35000
40000

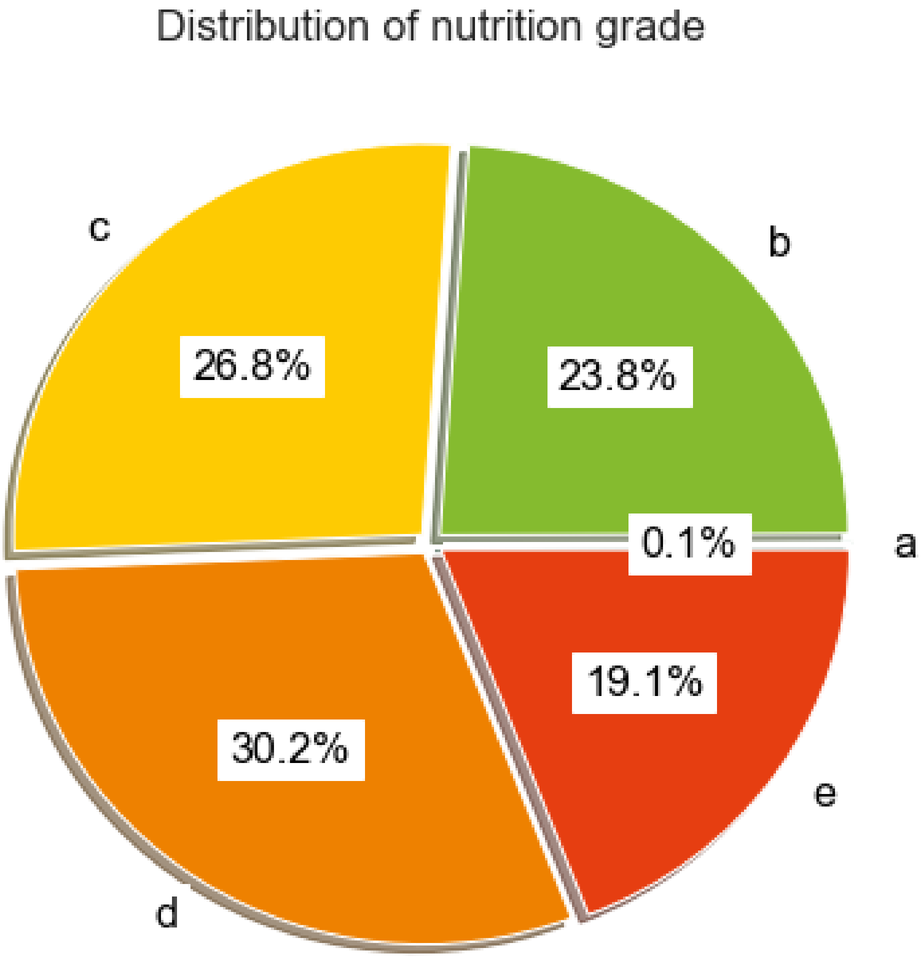
Boxplot

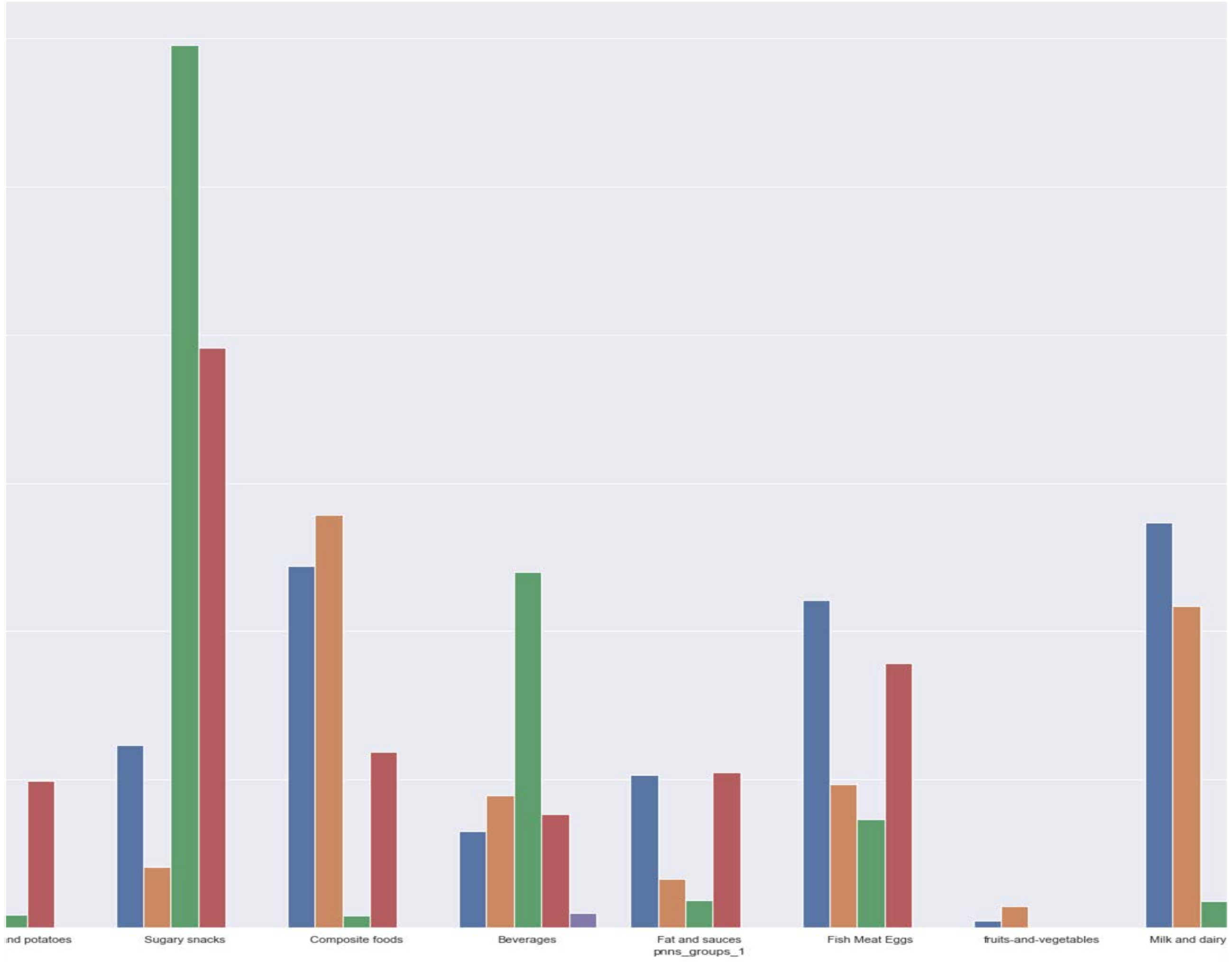

Numerical features Boxplot
0
500
1000
1500
2000
2500
3000
3500
energy_100g
0
5
10
15
20
25
proteins_100g
0.0
0.5
1.0
1.5
2.0
2.5
3.0
salt_100g
0.0
0.2
0.4
0.6
0.8
1.0
1.2
sodium_100g
0
10
20
30
40
50
60
70
sugars_100g
0
10
20
30
40
50
fat_100g
0
20
40
60
80
100
carbohydrates_100g
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
saturated-fat_100g
Boxplot : additives_n
0
5
10
15
20
25
30
35
nutrition-score-fr_100g
0
2
4
6
8
additives_n

Histogram

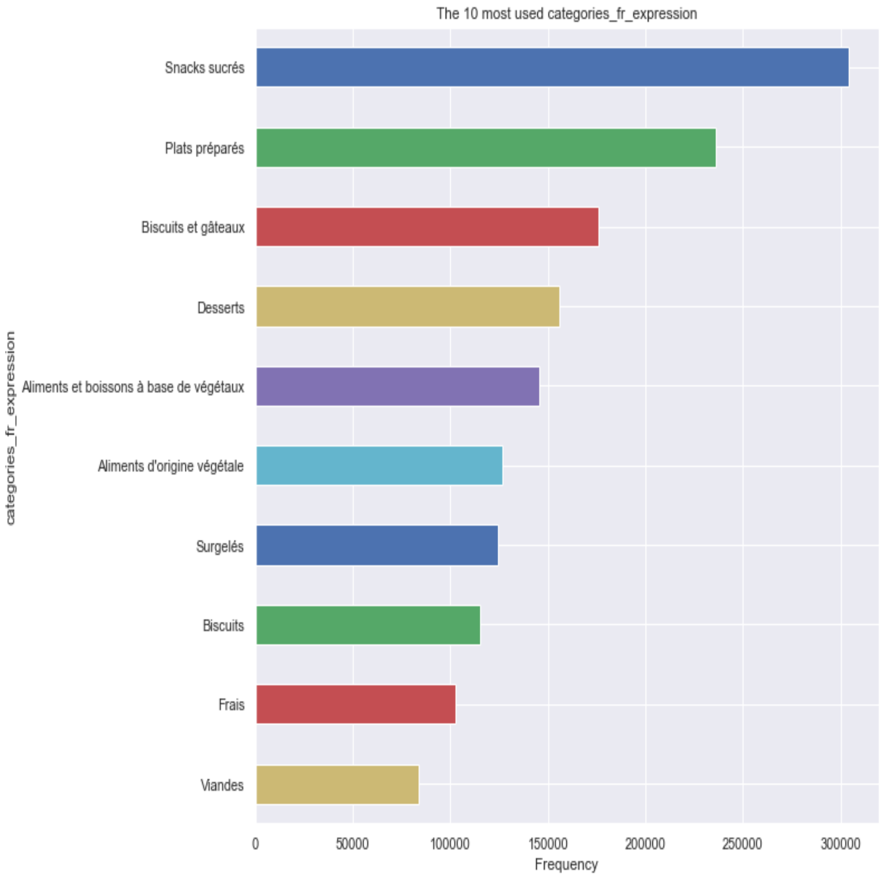

energy_100g
25000 20000 15000 10000 5000 0
0 1000 2000 3000
proteins_100g
30000 20000 10000 0
0 5 10 15 20 25
salt_100g
30000 20000 10000 0
0 1 2 3
sodium_100g
30000 20000 10000 0
0.00 0.25 0.50 0.75 1.00 1.25
ingredients_that_may_be_from_palm_oil_n
40000 30000 20000 10000 0
0 2 4 6 8
ingredients_from_palm_oil_n
40000 30000 20000 10000 0
0 2 4 6 8
additives_n
40000 30000 20000 10000 0
0 2 4 6 8
sugars_100g
40000 30000 20000 10000 0
0 20 40 60
fat_100g
40000 30000 20000 10000 0
0 20 40
carbohydrates_100g
25000 20000 15000 10000 5000 0
0 20 40 60 80 100
saturated-fat_100g
40000 30000 20000 10000 0
0 5 10 15 20
nutrition-score-fr_100g
30000 25000 20000 15000 10000 5000 0
0 10 20 30
nutrition-score-uk_100g
25000 20000 15000 10000 5000 0
0 10 20 30
fiber_100g
40000 30000 20000 10000 0
0 2 4 6 8
cholesterol_100g
60000 50000 40000 30000 20000 10000 0
0.00 0.02 0.04 0.06
trans-fat_100g
100000 80000 60000 40000 20000 0
−0.4 −0.2 0.0 0.2 0.4
calcium_100g
60000 50000 40000 30000 20000 10000 0
0.00 0.05 0.10 0.15 0.20
vitamin-c_100g
80000 60000 40000 20000 0
0.0000 0.0025 0.0050 0.0075 0.0100 0.0125
iron_100g
60000 40000 20000 0
0.000 0.001 0.002 0.003 0.004 0.005
vitamin-a_100g
60000 50000 40000 30000 20000 10000 0
0.00000 0.00005 0.00010 0.00015 0.00020

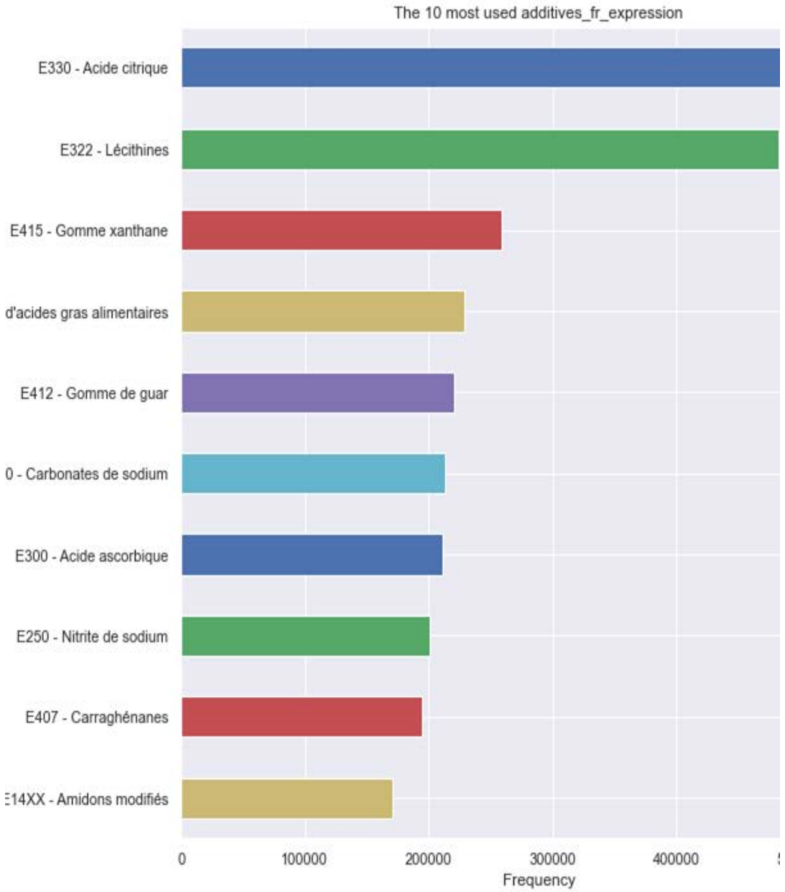
Categorical features visualisation

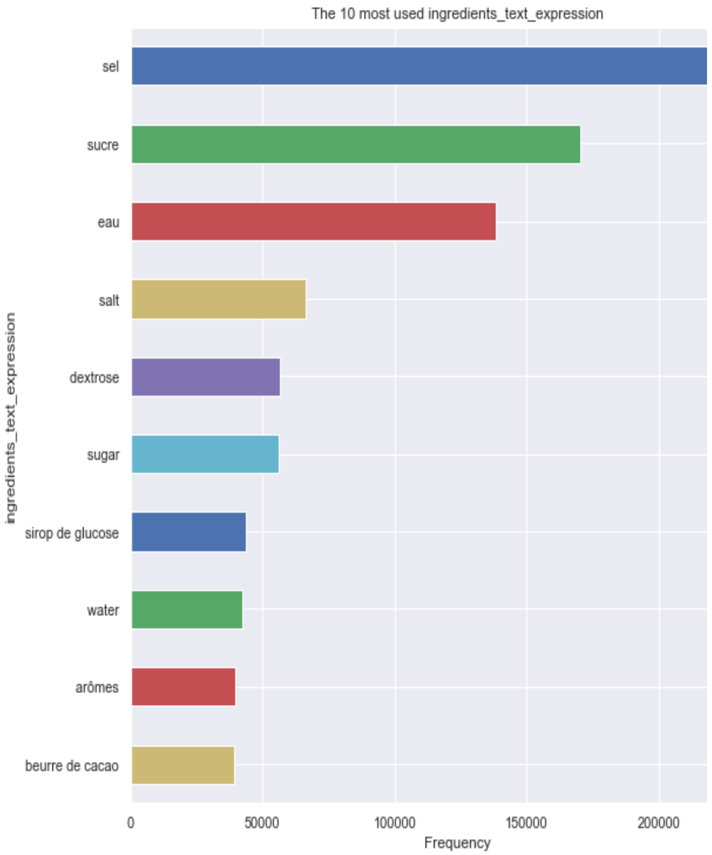




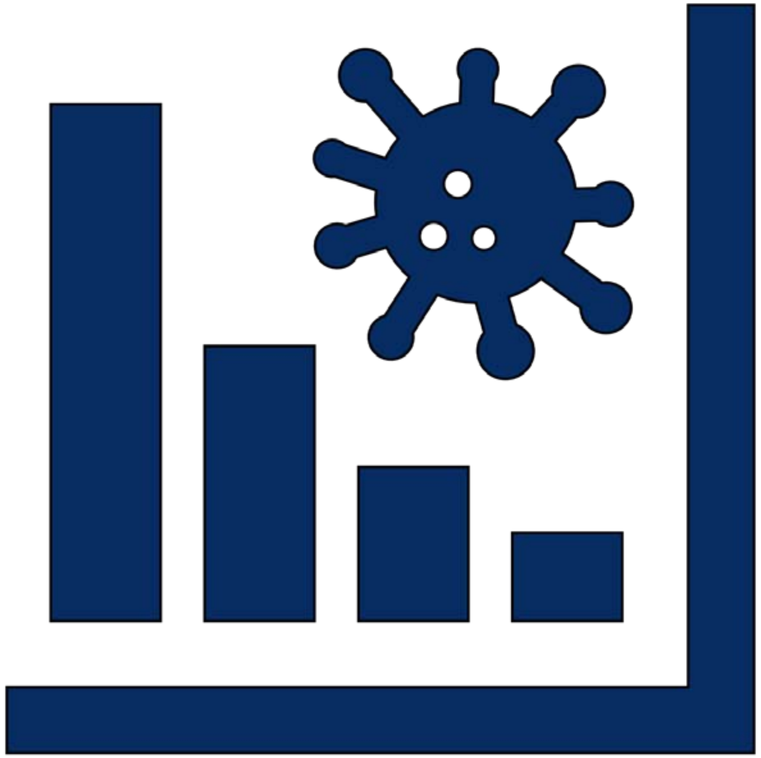
Categorical features visualisation



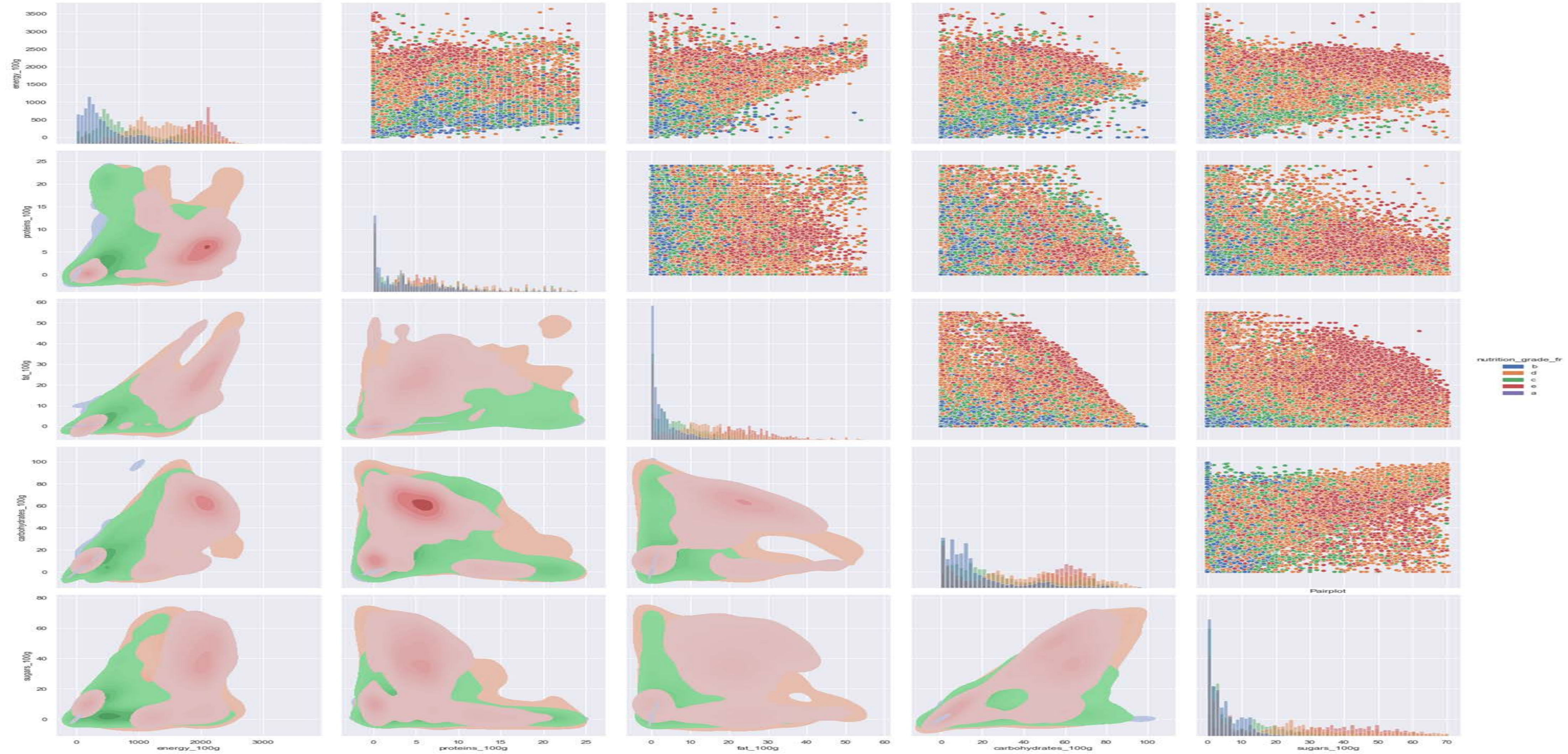




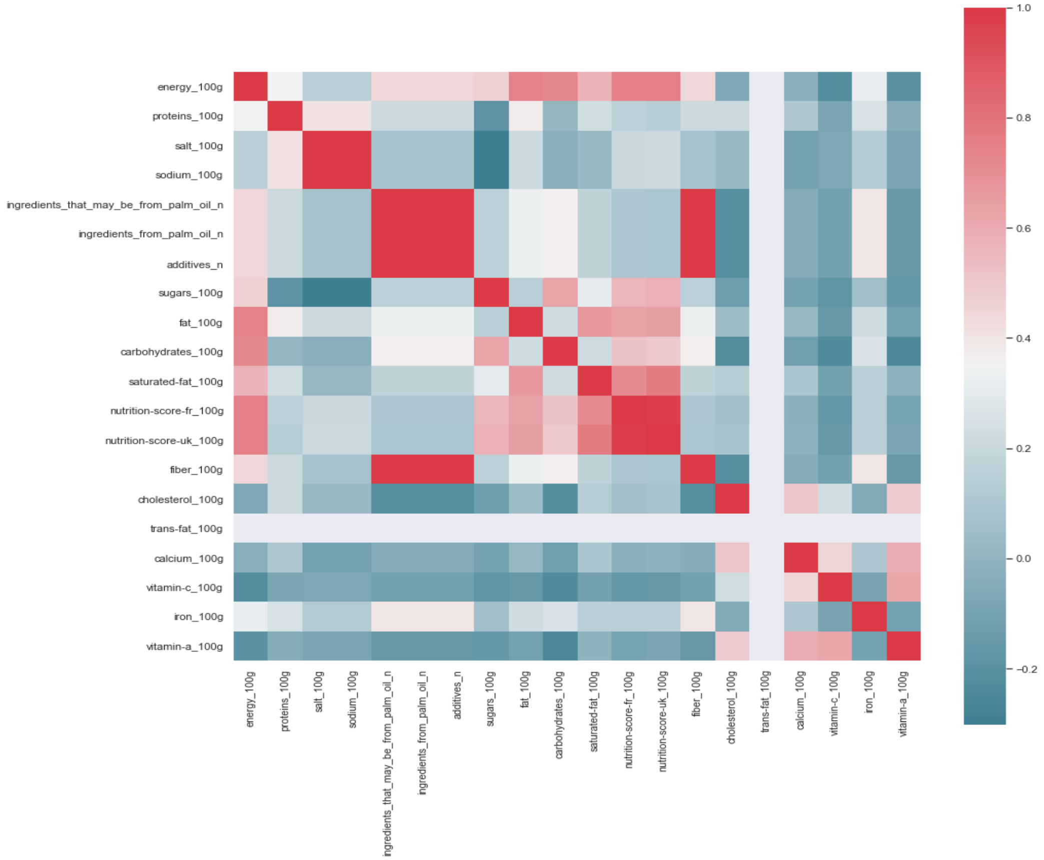
Bivariate analysis

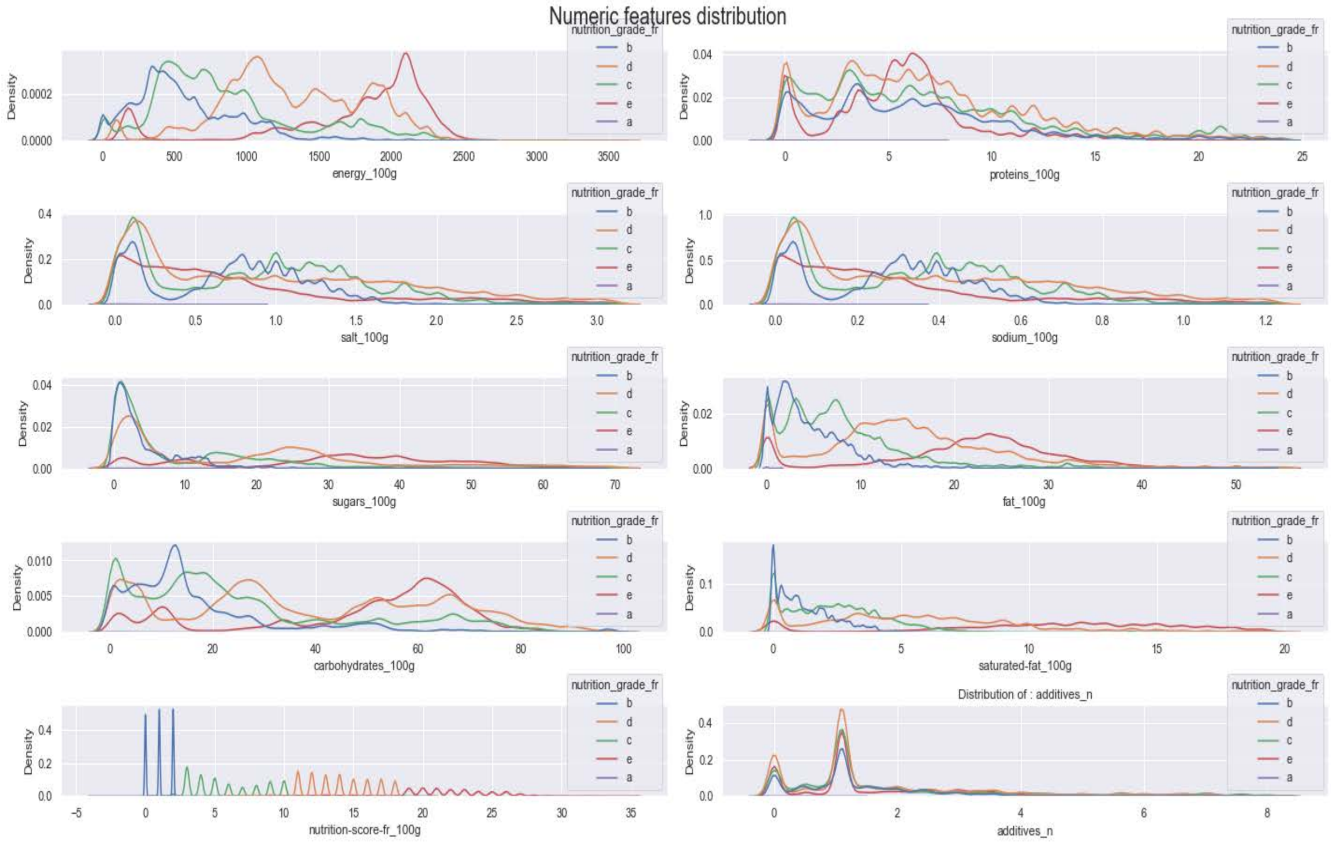


Features relationship



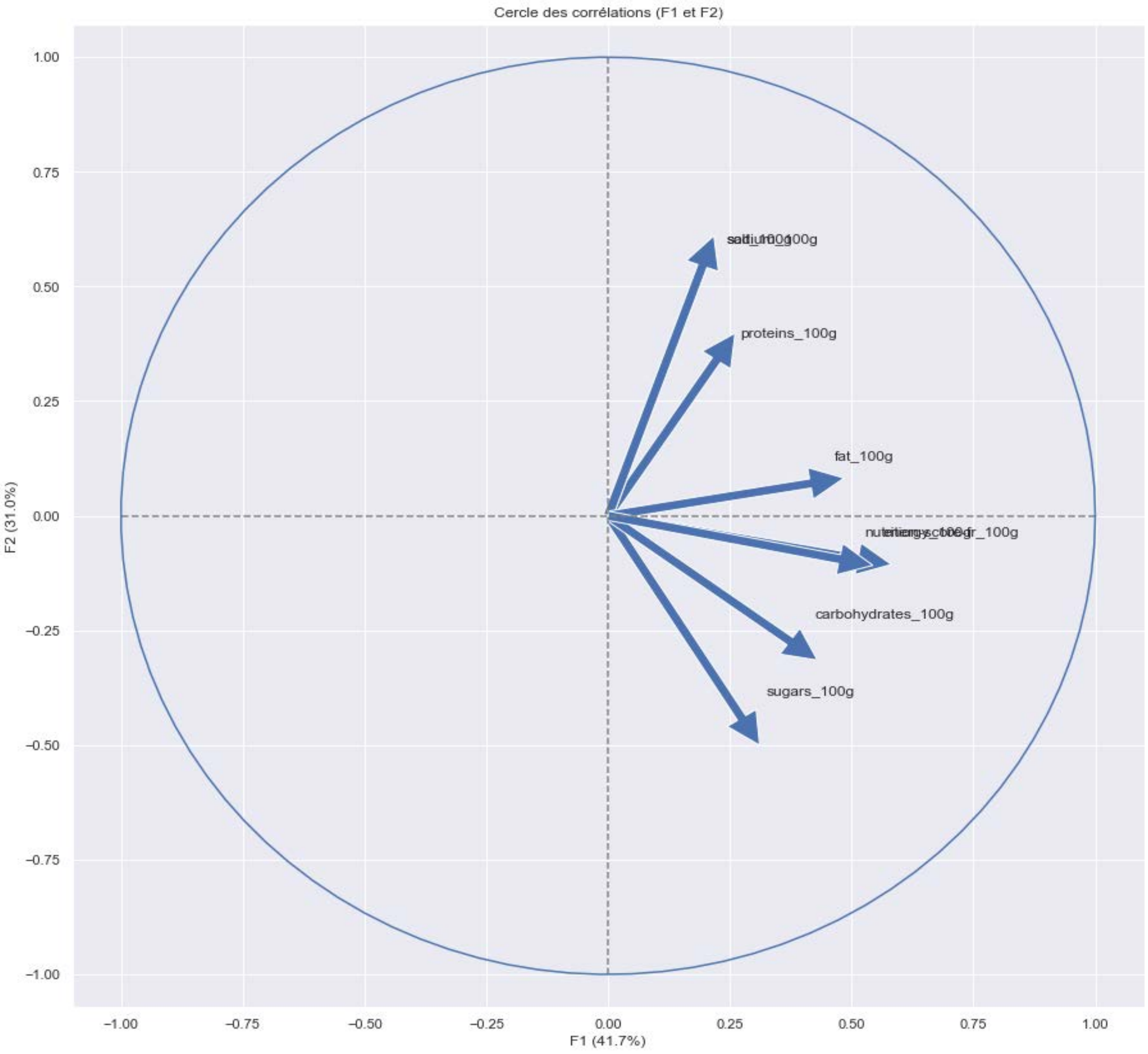
Correlation analysis

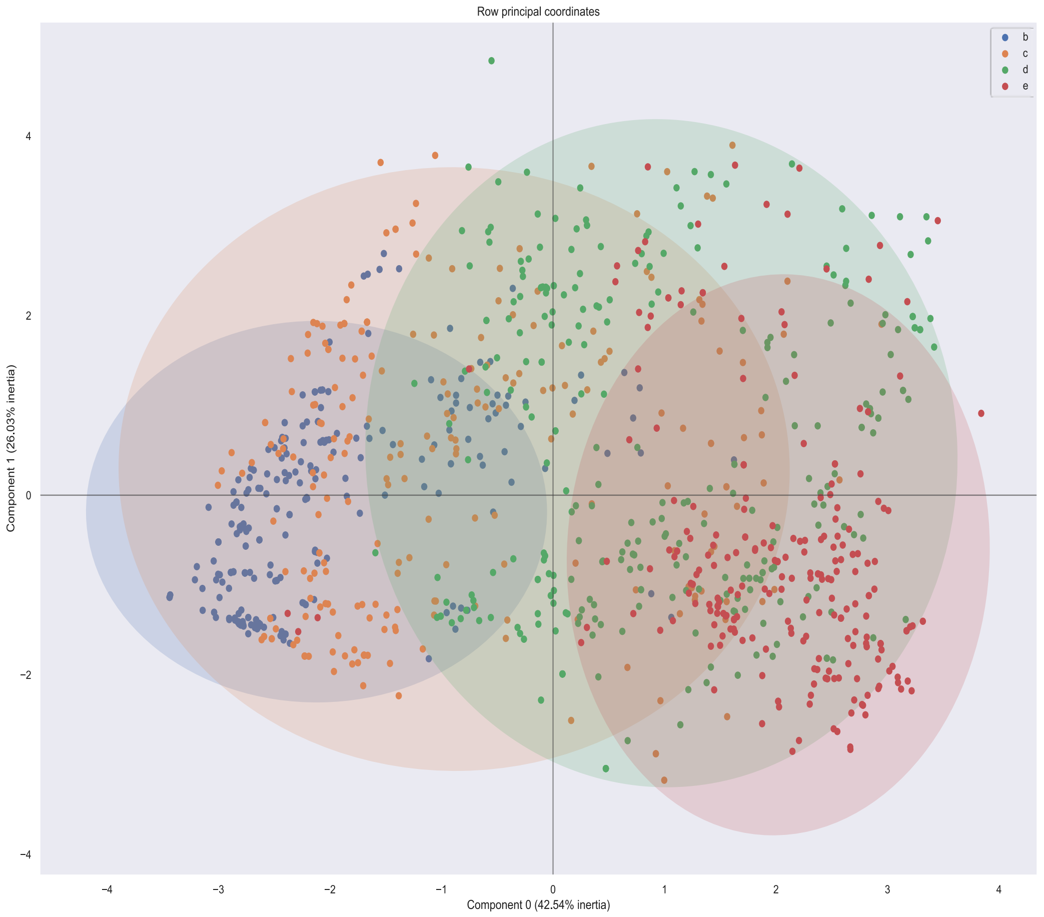




Multivariate analysis

PCA

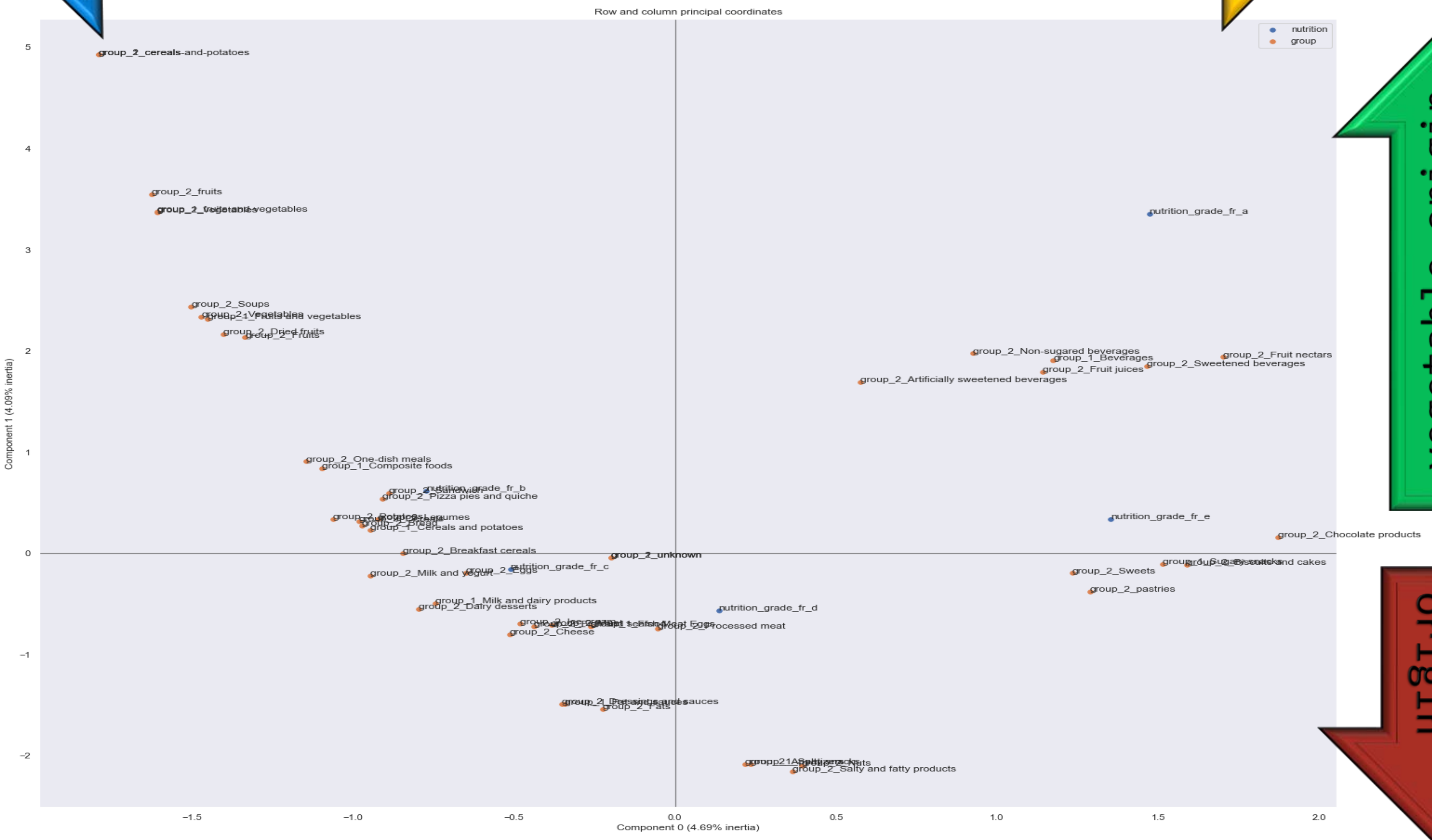

Cercle des corrélations (F1 et F2)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
F2 (31.0%)
sodium_100g
salt_100g
proteins_100g
fat_100g
nutrition-score-fr_100g
energy_100g
carbohydrates_100g
sugars_100g
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
F1 (41.7%)


Row principal coordinates
● b
● c
● d
● e
4
2
0
−2
−4
Component 1 (26.03% inertia)
−4
−3
−2
−1
0
1
2
3
4
Component 0 (42.54% inertia)

MCA

Healthy

Unhealthy



vegetable origin

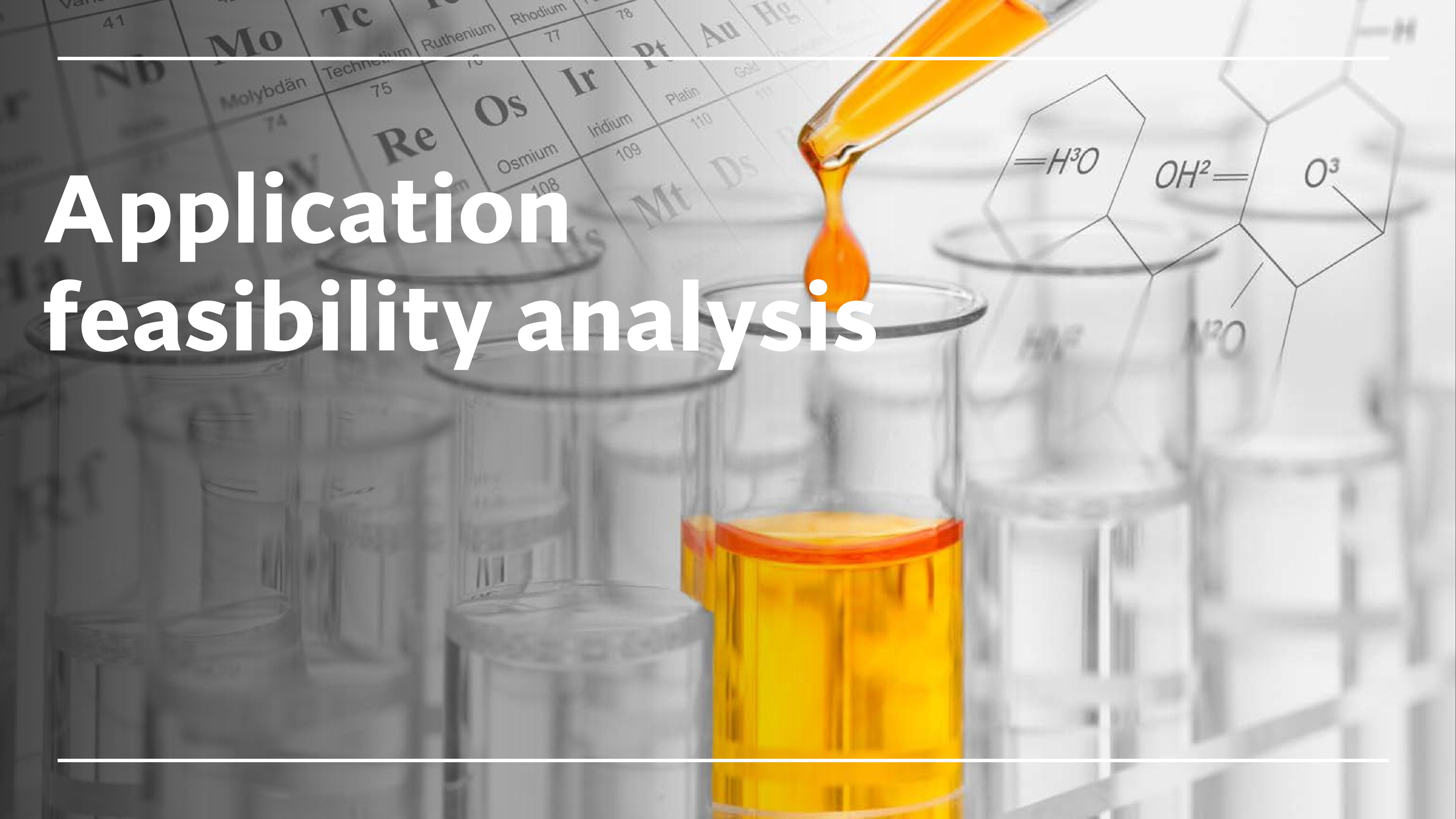
Animal origin

Statistic tests

	Chi-square	Degree of freedom	Critical value	p_value
pnns_groups_1	11171.42	44.	60.48	0.0
pnns_groups_2	18517.63	156	186.15	0.0
countries	5030.93	2180	2289.74	$5.92.10^{-226}$
categories_fr	77966.73	34084	34514.59	0.0
additives_fr	104693.14	63472	64059.18	0.0

	df	sum_sq	mean_sq	F	PR(>F)
pnns_groups_1	11.0	5.48e+05	49809.45	11308.08	0.0
nutrition_grade_fr	4.0	1.06e+06	263771.33	59883.17	0.0
pnns_groups_1:nutrition_grade_fr	44.0	2.07e+04	469.62	106.62	0.0
Residual	29297.0	1.29e+05	4.40	NaN	NaN

- Our features not satisfy normality test



Application feasibility analysis

Application feasibility analysis

- Data availability
- Relationship between categorical features and healthy diet
- Relationship between features
- Statistical test
- Attention!
  - Data reporting
  - Nan filling

Thank You for your attention!