Feasibility Study of an application for public health

JANUARY 2022

NAJWA MOULINE

Presentation Outline

1. My application idea
2. Dataset Cleaning
3. Dataset Exploratory Analysis
4. Presentation of facts relevant to the application

My application idea

Context

Context

- The agency is launching a call for projects to find innovative ideas for applications related to food.
- Open Food Fact dataset available

Business Problem

- Answer the call and identify an application idea

Mission

- Consult the dataset provided to conclude if it allows to realise the idea, in response to the call.

Methodology

1. Process the dataset
2. Perform univariate analysis of relevant variables
3. Perform multivariate analysis of relevant variables
4. Justify the idea of application and conclude its feasibility

My application idea

The idea

Application NutrINF

- Nutrition of children under 3 years of age
- Operation:

Entry of foods and associated servings consumed/to be consumed and age

Nutrition of children under 3 years of age

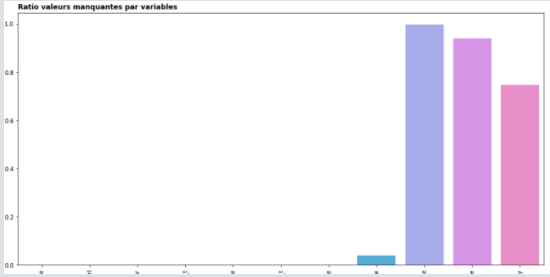
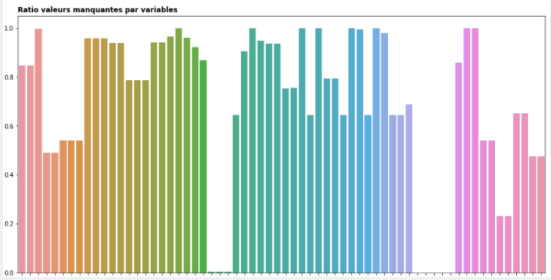
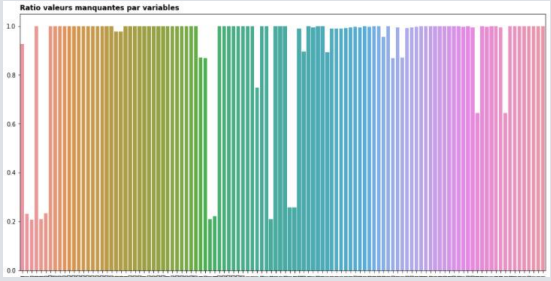
- Based on the amount of servings in 5 food categories

Category (portion size)	Minimum servings / day	
	1 to 2 years	2 to 3 years
Legumes and legumes (75g)	2-3	2 ½
Fruits (350kJ)	½	1
Grain and grain foods (500kJ)	4	4
Lean meat and poultry, fish, eggs, tofu, nuts and seeds, and legumes (550kJ)	1	1
Milk, cheese with yogurt and/or substitutes (550kJ)	1-1 ½	1 ½
Other (considered by default nutrigrade non A)	0	0

Dataset Cleaning

Description of the dataset

File description

Data (2054909 obs. x 187 columns)				
Product Information				
Section	Top 5 most filled variables	Type variables	% NaN	Ratio NaN per variable
Variables _ general information	Code, url, creator, created_datetime , product_name	Object / int64 / datetime	25	
Variables _ set of tags & ingredients	Countries, pnns_groups_1, pnns_groups_2, states, images	Object / Float	69	
Variables - nutritional information	Energy_kj_100g, energy_100g, Proteins_100g Fat_100g, Carb_100g	Float	92	

Dataset Cleaning

Cleaning operations

Cleaning process followed in 4 steps

Simplification – filter data to:

- Products sold in France
- Products without duplicates

Selecting variables – removing:

- Variables not filled in
- Redundant variables
- Non-relevant variables

Identifying outliers – removing :

- Non-logical values
- Impossible values

Handling missing values – filling according to the variable by different methods:

- Values at 0
- Average values by category
- Calculated values
- Imputed values by classification and backward

Dataset Cleaning

Simplification and relevance of the dataset

Variables : 187

Products : 562410

NaN : 78%

Simplifying the dataset

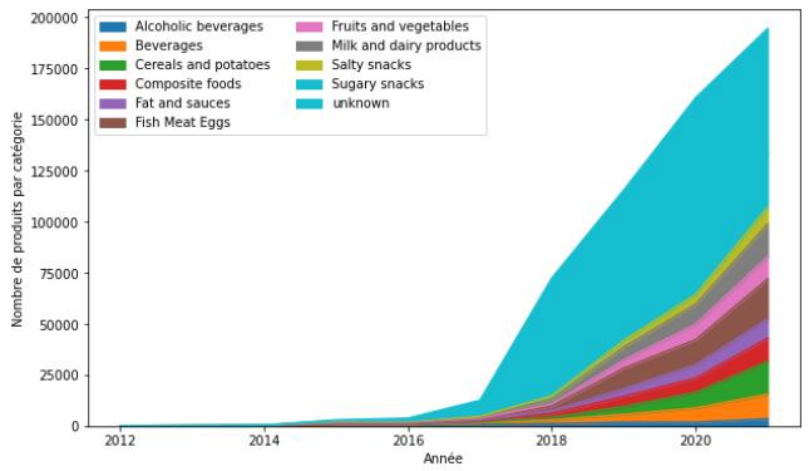
Initial dataset	
Nb of products: 2054909	Fill rate: 20%

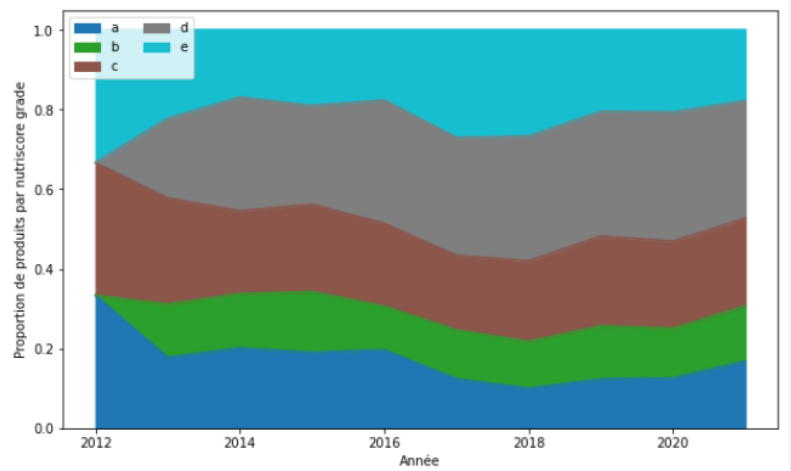
Filter on products sold in France (countries includes FR)	
Nb of products: 562417	Fill rate: 22%

Filter on non-redundant products (unique code, incorrectly filled in duplicates or "old" deleted)	
Nb of products: 562410	Fill rate: 22%

Relevance of the simplified dataset for the application of the idea

The dataset completed over a period mainly of 5 years (2016-2021), and including a number of products in each PNNS category and for each nutrigrade during this period.





Dataset Cleaning

Selecting variables

Variables : 30

Products : 562410

NaN : 58%

Selecting variables

Simplified dataset

Nb variables: 187	Fill rate: 22%

1- Variables filled in (at least one non-NaN over the entire variable)

Nb variables: 147	Fill rate: 28%

2 – Variables relevant to the idea (name, category, nutriscore, nutritional var.)

Nb variables: 48	Fill rate: 36%

3 - Removal of redundant variables (and filling in the selected variable)

Nb variables: 30	Fill rate: 42%

#	Column	Non-Null Count	Dtype
0	code	562410 non-null	object
1	product_name	542074 non-null	object
2	pnns_groups_1	562395 non-null	object
3	nutriscore_grade	269454 non-null	object
4	nutrition_score	269457 non-null	float64
5	energy_100g	460177 non-null	float64
6	fat_100g	457355 non-null	float64
7	saturated_fat_100g	457415 non-null	float64
8	monounsaturated_fat_100g	3045 non-null	float64
9	polyunsaturated_fat_100g	3066 non-null	float64
10	omega_3_fat_100g	1390 non-null	float64
11	omega_6_fat_100g	302 non-null	float64
12	omega_9_fat_100g	54 non-null	float64
13	trans_fat_100g	2912 non-null	float64
14	cholesterol_100g	2933 non-null	float64
15	carbohydrates_100g	457278 non-null	float64
16	sugars_100g	458484 non-null	float64
17	starch_100g	298 non-null	float64
18	polyols_100g	762 non-null	float64
19	fiber_100g	143338 non-null	float64
20	soluble_fiber_100g	152 non-null	float64
21	insoluble_fiber_100g	150 non-null	float64
22	proteins_100g	459167 non-null	float64
23	casein_100g	49 non-null	float64
24	serum_proteins_100g	44 non-null	float64
25	nucleotides_100g	21 non-null	float64
26	salt_100g	448118 non-null	float64
27	sodium_100g	448117 non-null	float64
28	fruits_vegetables_nuts_100g	250892 non-null	float64
29	pnns_groups_12	562395 non-null	object

Dataset Cleaning

Removing outliers

Variables : 30

Products : 562410

NaN : 60%

Identification of outliers in nutriscore/nutrigrade variables

1 - Assigning NaN value to values if Nutriscore does not have an integer value between -15 and 40.

Nutriscore

Nutrigrade

Nutriscore

2 - Assigning NaN value to values if Nutrigrade does not have one of these 5 values: A, B, C, D or E

Identification of outliers in nutritional variables

3 - Assign NaN value to values if ∑ (Sat Fat + Other Fat) > Fat

4 - Assigning NaN value to values if ∑ (Sugar + Fibres + Other carbs) > Carbohydrate

Sat Fat | Other Fat | Sugars | Fibres | Other carbs | Other prot

Fat | Carbohydrate | Protein

5 - Assigning NaN value to values if ∑Other prot > Protein

Energy | Salt | Fruit Veg

6 - Assigning NaN value to all energy values outside of 0-3500kJ
7 – Assinging NaN value to all nutritional variables on 100g apart from 0-100g

Dataset Cleaning

Handling missing values

Variables : 15

Products : 562410

NaN : 4%

Different treatments of missing values by variable

Sat Fat | Other Fat | Sugars | Fibres | Other carbs

1 - Assignment value 0

Fat | Carbohydrate | Protein

2 - Assignment value Average of the variable by category PNNS1

Energy

Energy (kJ) = 17 * Carbohydrates (kJ) + 37 * Fat (kJ) + 17 * Protein (kJ)

3 - Assignment value calculated by the formula: energy kJ = 37*fatkJ + 17*carbkJ + 17*proteinkJ

Salt | Sat Fat | Sugars | Protein | Fibres | Fruit Veg

Nutriscore

(source Santepubliquefrance.fr) The logo is awarded on the basis of a score taking into account per 100g or 100mL of product, the content:
- In nutrients and foods to promote (fiber, proteins, fruits, vegetables)
- In nutrients to limit (energy, saturated fatty acids, sugars and salt)

4 - Assigning the value of the nutriscore/nutrigrade in backward fill mode (limited to 20) on the dataframe classified by dependent variables and by order of filling of the variable.

Dataset Cleaning

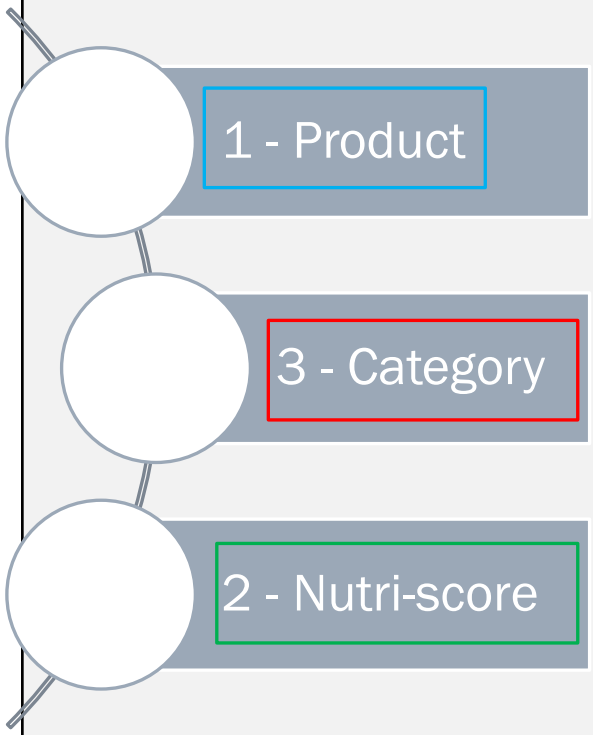
Cleaned dataset

Variables : 15

Products : 562410

NaN : 4%

Cleansed dataset



Product / Category

Variable	Nb mod.	Type	% NaN
Code	51514	Object	0
Product_name	39394	Object	3,6
pnns_groups_1	11	Object	0
pnns_groups_12	14	Object	0

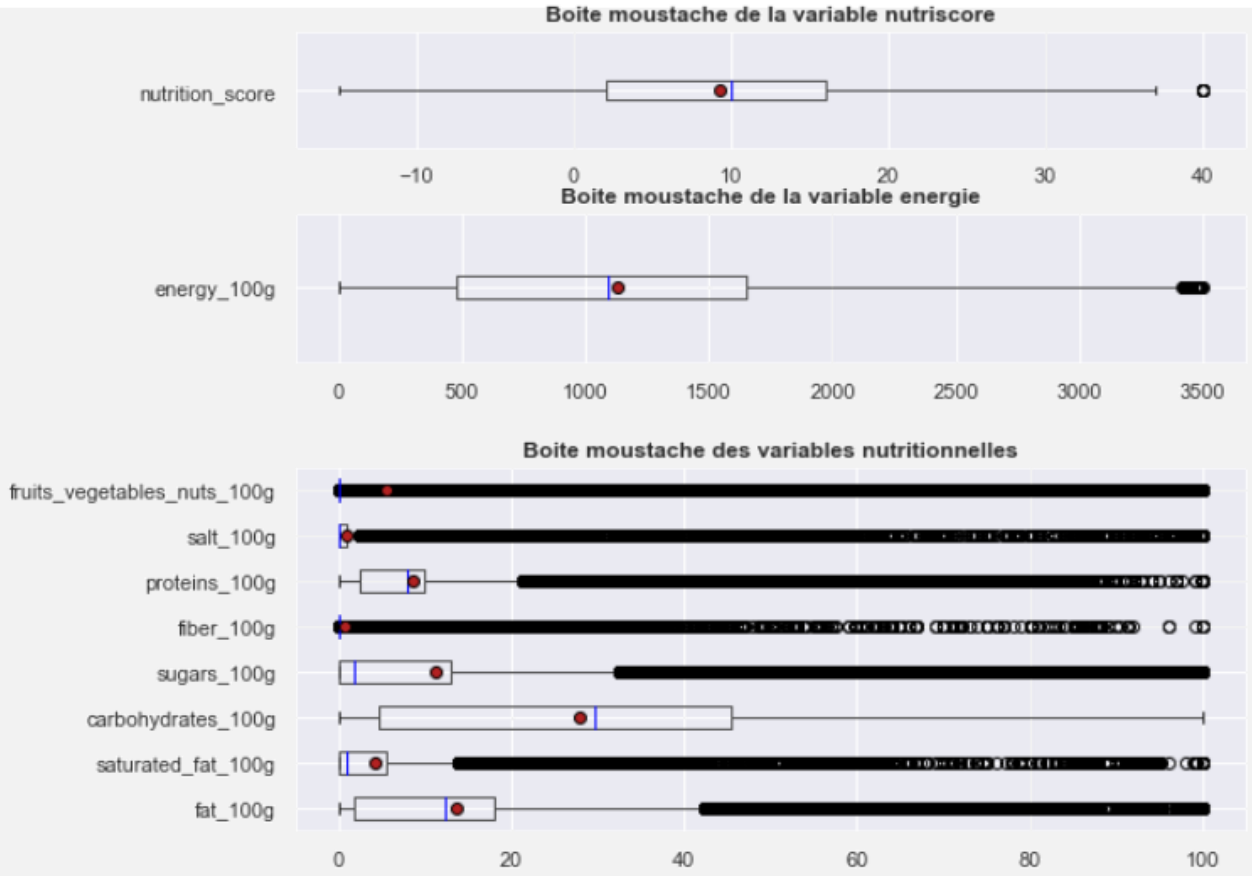
Nutri-score

Variable	Values	Type	% NaN
nutriscore_score	-15 - 40	Float	18,1
nutriscore_grade	A, b, c, d, e	Float	18,1
energy_100g	0-3500	Float	20,9
fat_100g	0-100	Float	0
saturated-fat_100g	0-100	Float	0
proteins_100g	0-100	Float	0
carbohydrates_100g	0-100	Float	0
sugars_100g	0-100	Float	0
fiber_100g	0-100	Float	0
salt_100g	0-100	Float	0
fruits_vegetables_nuts_100g	0-100	Float	0

Dataset Exploratory Analysis

Univariate analysis of quantitative variables (1 of 2)

Boxplot of quantitative variables



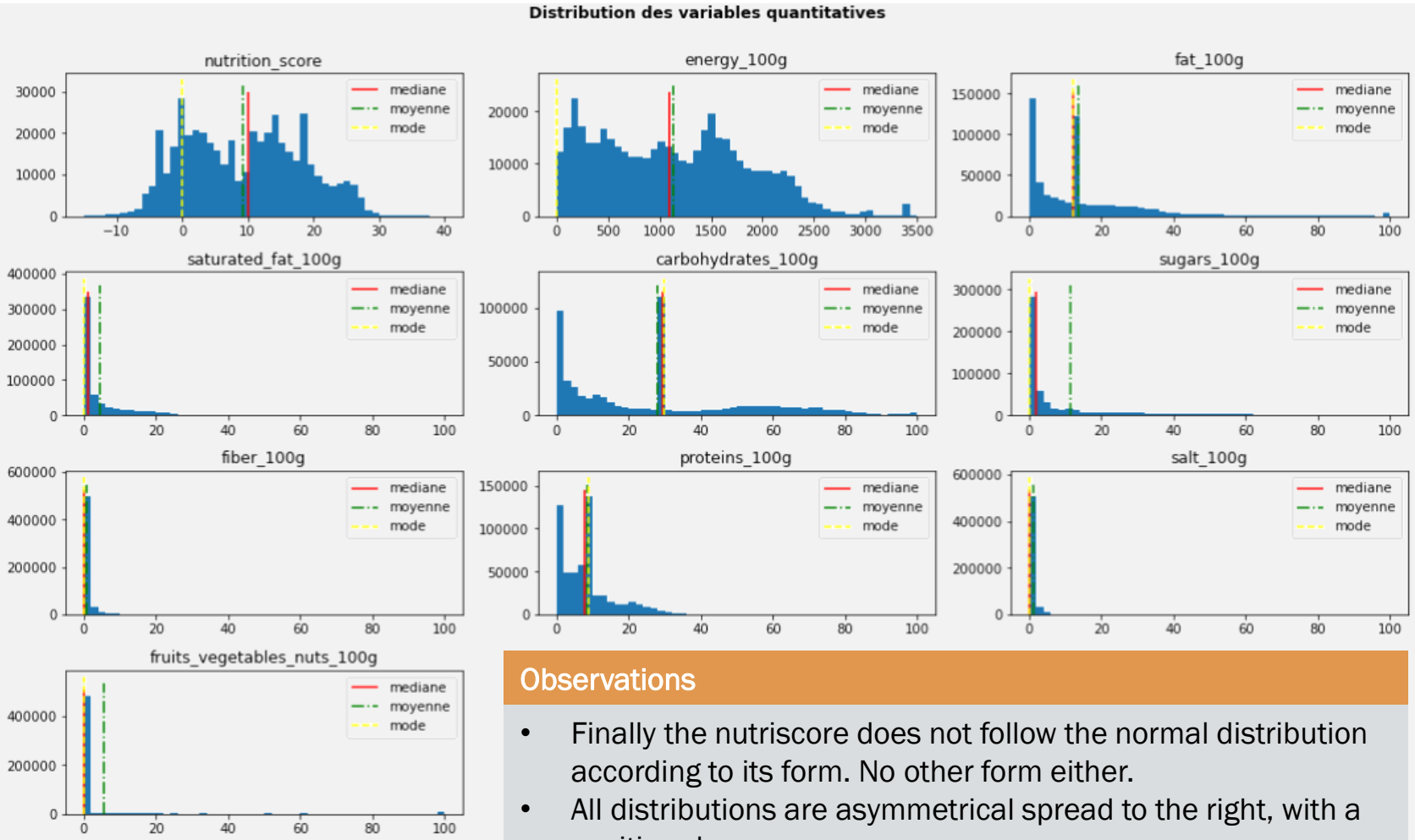
Observations

- The nutriscore seems to have an almost normal distribution.
- The distributions of the other variables are very different and different from each other. None of their distribution seems to follow a normal distribution.
- Some are very dispersed – carbohydrates. Others very tight – proteins
- => The Nutriscore does not appear to have a direct link to the other variables.

Dataset Exploratory Analysis

Univariate analysis of quantitative variables (2 of 2)

Histogram of quantitative variables



Observations

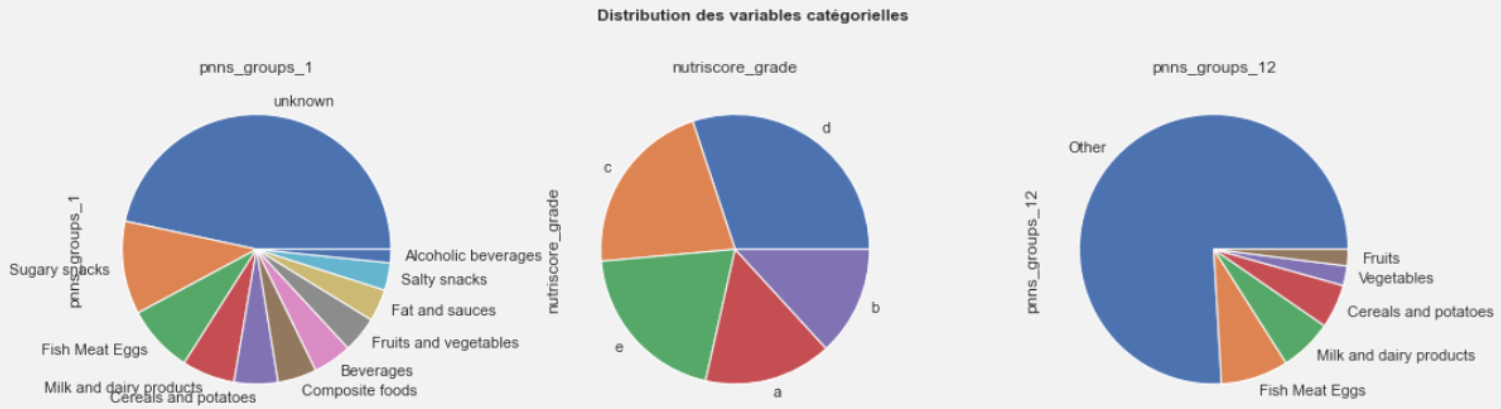
- Finally the nutriscore does not follow the normal distribution according to its form. No other form either.
- All distributions are asymmetrical spread to the right, with a positive skewness.

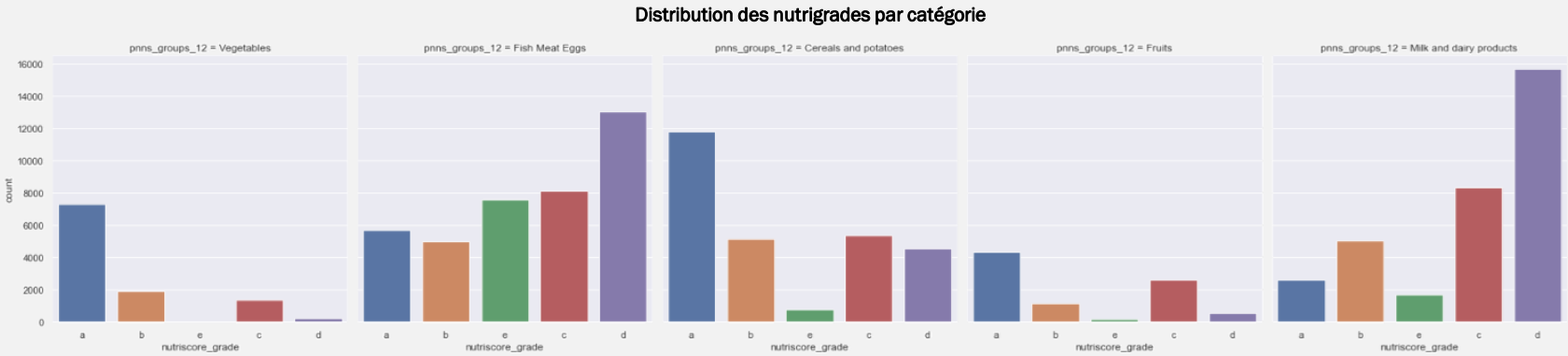
- Some distributions are concentrated - salt, fiber, fruit. Others less concentrated – nutriscore, energy.
- Some distributions are bimodal – nutriscore, energy and the main nutritional variables.
- => The Nutriscore has a distribution that has some similarities with some other variables.

Dataset Exploratory Analysis

Univariate and bivariate analysis of qualitative variables

Piechart and distribution of qualitative variables





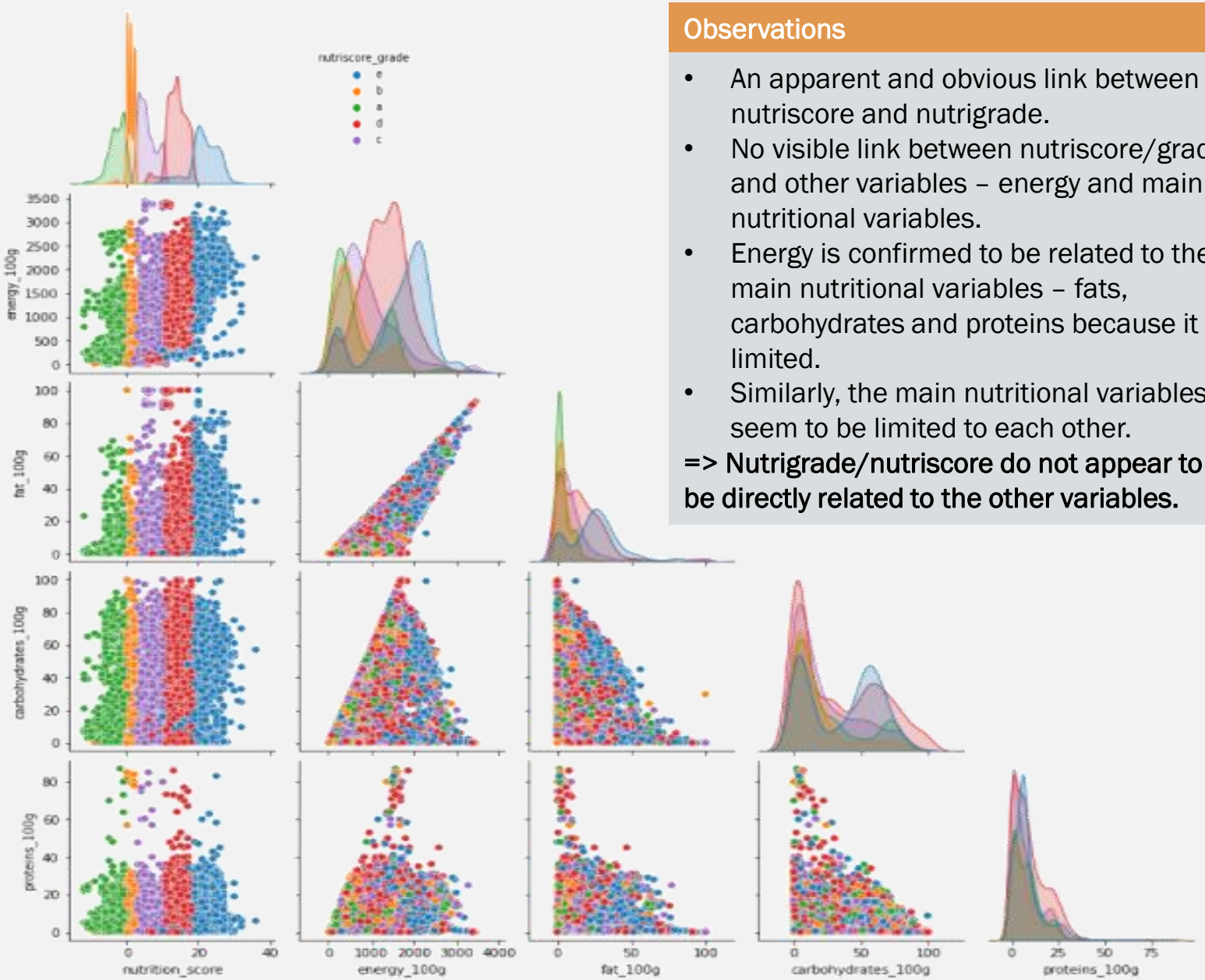
Observations

- The modalities of each variable generally have an equivalent proportion.
- The nutrigrade by category PNNS12 (categories that interest us) does not allow to conclude a precise relationship between them. Note that we have many products in each of Nutrigrade A.
- Nutrigrade A is dominant for vegetables/starchy foods/fruits. While nutrigrade E dominates the meat and dairy categories.
- => Nutrigrade varies from one category to another, hence a possible dependence.

Dataset Exploratory Analysis

Bivariate analysis of quantitative variables (1 of 2)

Pairplot of the main quantitative variables



Observations

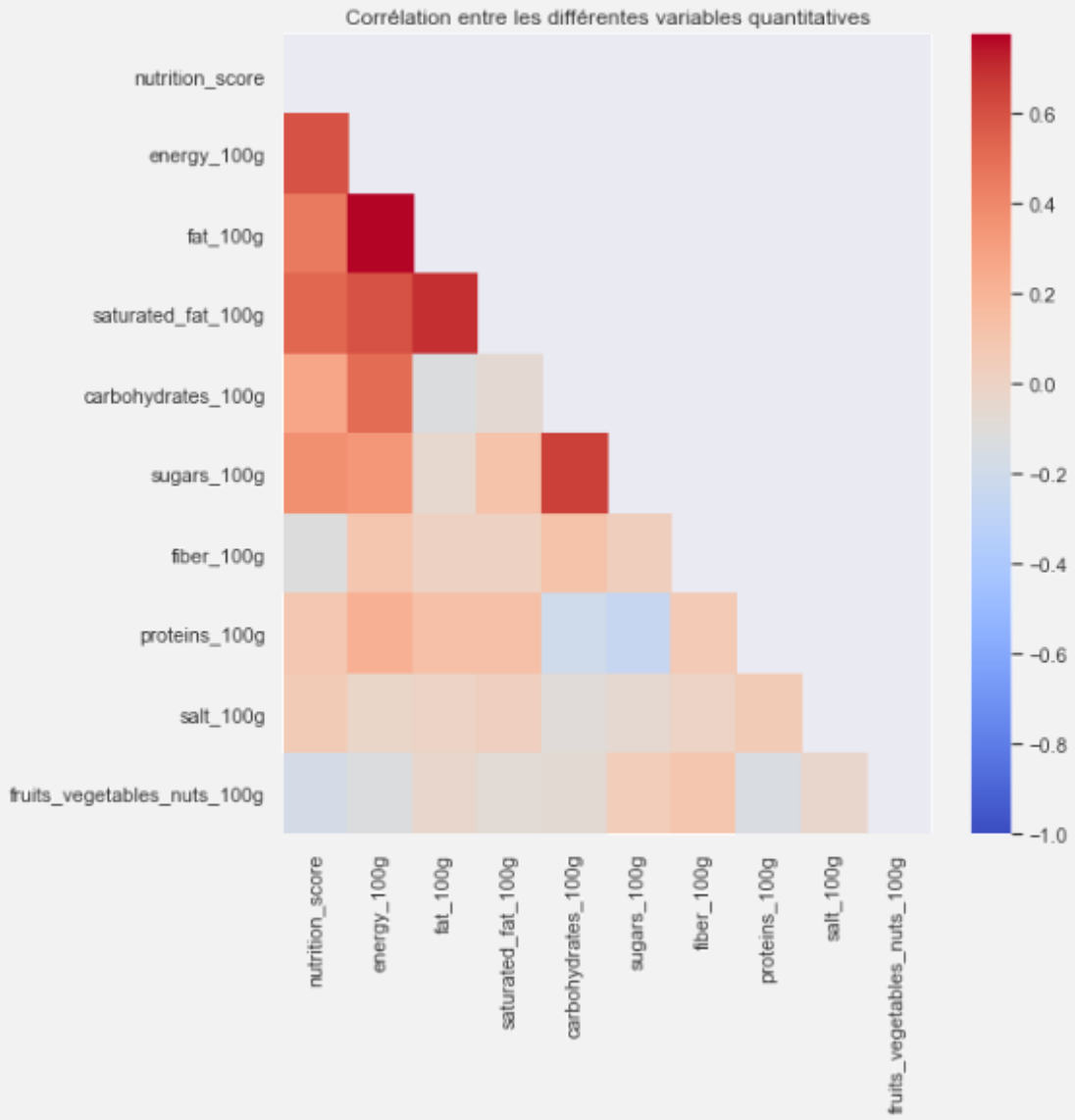
- An apparent and obvious link between nutriscore and nutrigrade.
- No visible link between nutriscore/grade and other variables – energy and main nutritional variables.
- Energy is confirmed to be related to the main nutritional variables – fats, carbohydrates and proteins because it is limited.
- Similarly, the main nutritional variables seem to be limited to each other.

=> Nutrigrade/nutriscore do not appear to be directly related to the other variables.

Dataset Exploratory Analysis

Bivariate analysis of quantitative variables (2 of 2)

Heatmap between quantitative variables



Observations

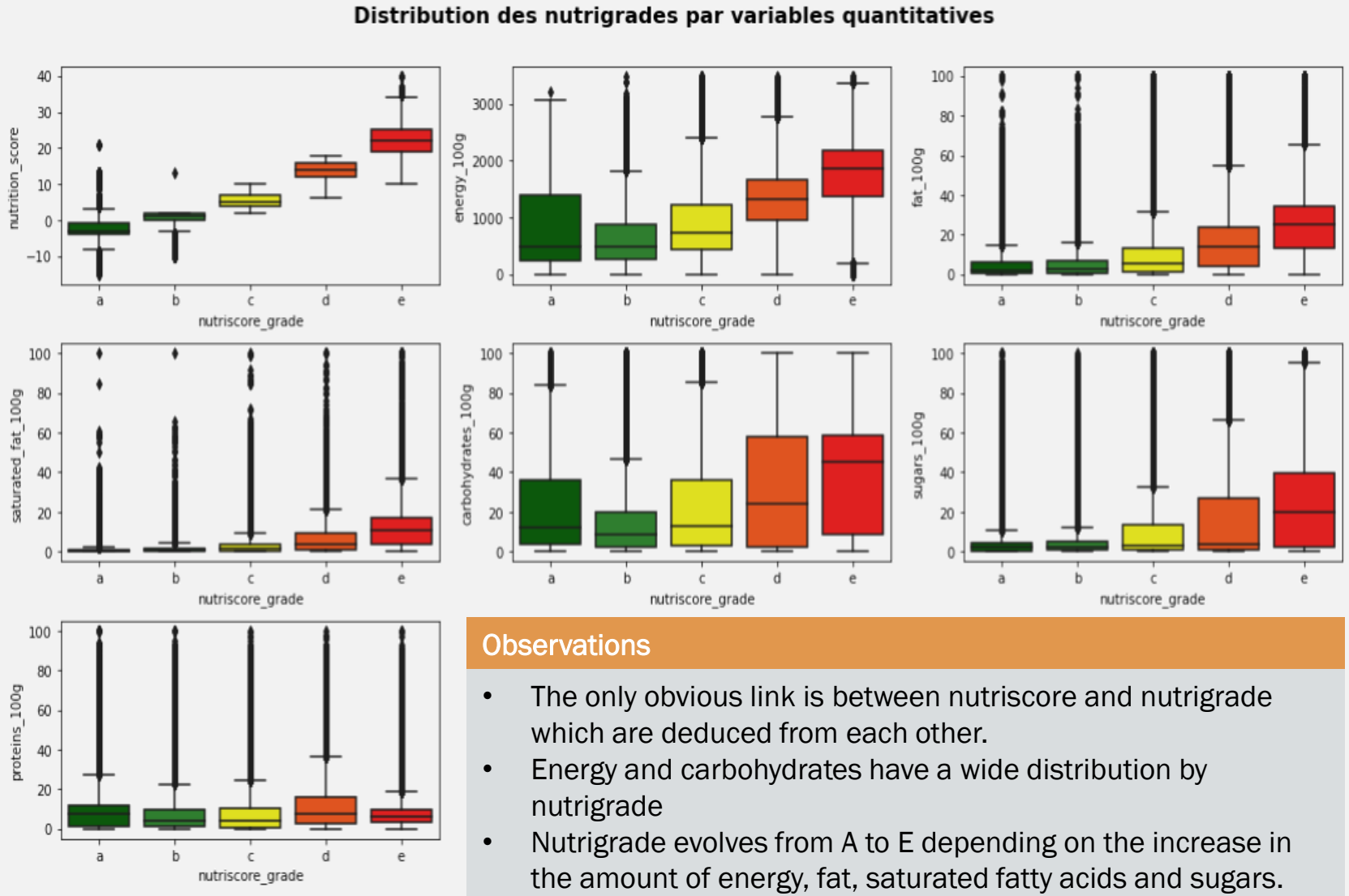
- The heatmap makes it possible to identify the dependence between the nutriscore and the nutritional variables.
- Nutriscore is highly correlated with energy, saturated fatty acids and fats, correlated with sugars and carbohydrates, and less with protein to have none with fiber and fruits/vegetables. This seems consistent.
- Energy has a strong correlation with fat, saturated fat and carbohydrates.

=> Nutriscore is dependent on other variables and more particularly on energy, saturated fatty acids, and lipids.

Dataset Exploratory Analysis

Bivariate analysis of qualitative and quantitative variables

Distribution of nutrigrade by quantitative variable (ANOVA)



Observations

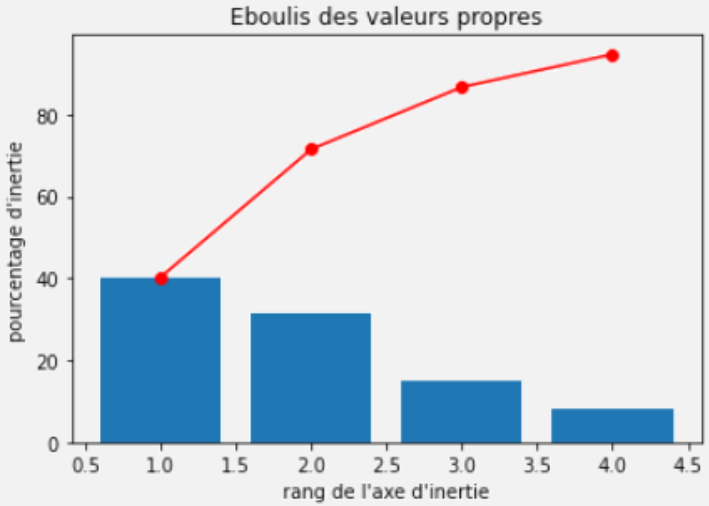
- The only obvious link is between nutriscore and nutrigrade which are deduced from each other.
- Energy and carbohydrates have a wide distribution by nutrigrade
- Nutrigrade evolves from A to E depending on the increase in the amount of energy, fat, saturated fatty acids and sugars.
- Proteins do not seem to have any impact on nutrigrade.

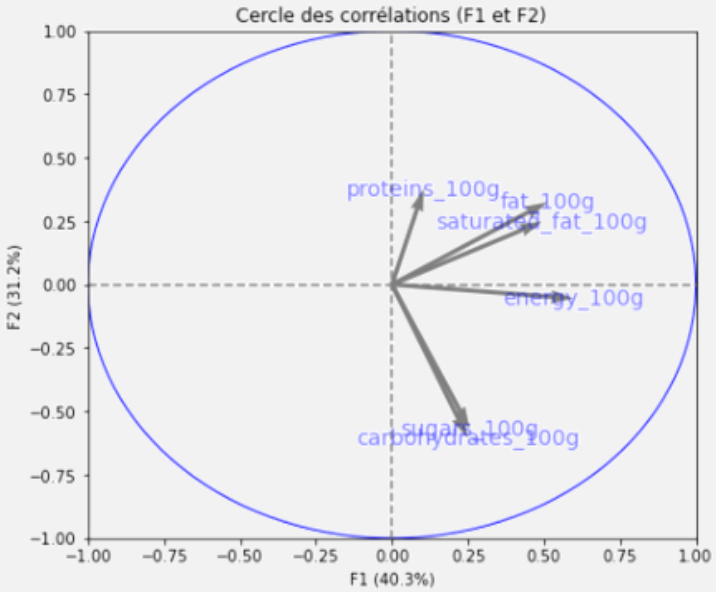
=> The Nutriscore confirms that it is highly dependent on energy, saturated fatty acids, fats, and sugars.

Dataset Exploratory Analysis

Multivariate analysis in main components

Principal Component Analysis

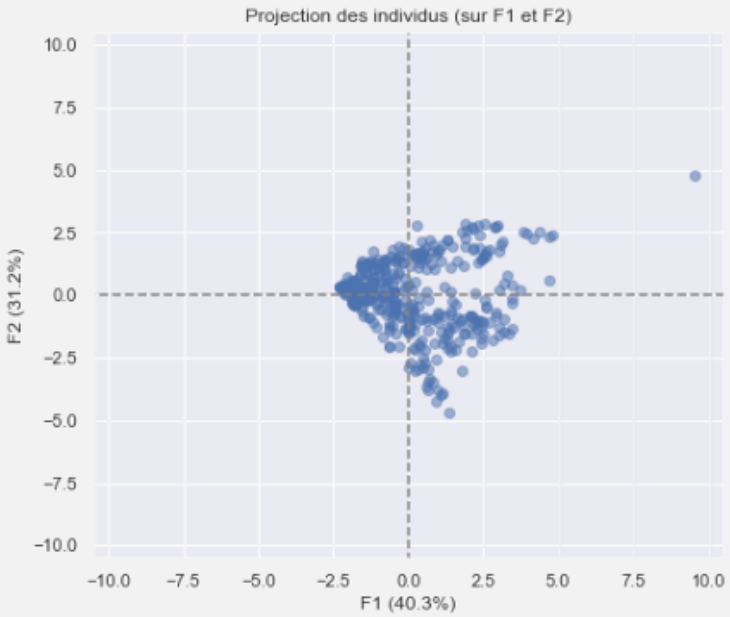




Observations

- The eigenvalues decomposition suggest the first 2 most relevant components (drop to axis 3)
- F1 and F2 account for 71% of total inertia.
- On F1, energy, fat and saturated fat are positively correlated. All variables are positive on this axis. This axis represents the amount of energy.
- On F2, carbohydrates and sugars are negatively correlated, unlike proteins and fats. This axis represents natural energy.
- The projection of products on these two axes is concentrated.

=> The PCA allows a reduction to 2 components.



Presentation of facts relevant to the application

Observations

Observations of the dataset exploratory analysis

Observations

- Dependency between variables.
- Strong correlation between nutriscore and certain variables such as energy, saturated fatty acids and lipids.
- Reduction of 6 initial variables to 2 main components representing 71% of the total inertia.

Presentation of facts relevant to the application

Conclusions

Relevance and limitation of the dataset

Reminder – the needs for the idea

- The idea requires product information, 5 specific categories and nutriscore/nutrigrade (especially nutrigrade A).

Feasibility of the idea

- The database is regularly filled and contains more than 500,000 products in France spread over all categories and nutrigrade.
- The nutriscore /nutrigrade of the products not known can be deduced from the two new compound variables 1 and 2. Compound variables can be inferred for 79% of products.
- The product category is known for about 70% of the products.
- The idea is feasible but limited to products with categories and for which component variables are provided.

Thank you for your attention!