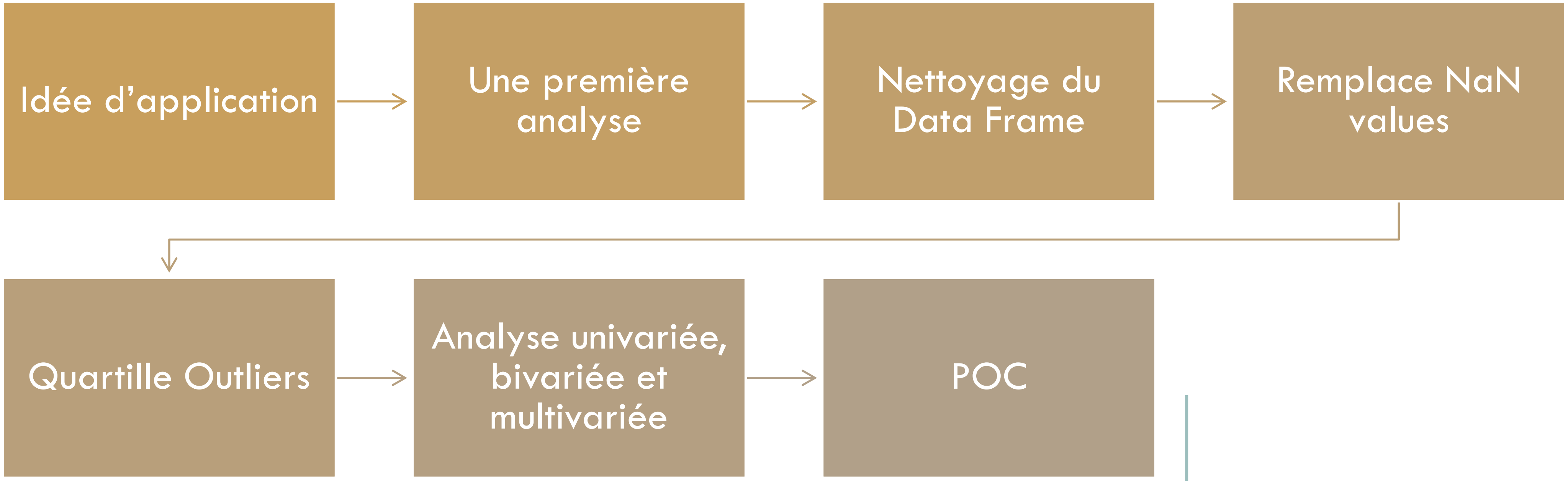
open FOOD facts

CONCEVEZ UNE APPLICATION AU SERVICE DE LA SANTÉ POUBLIQUE

Emanuele Partenza

CONCEVEZ UNE APPLICATION AU SERVICE DE LA SANTÉ PUBLIQUE:



Idée d'application

Une première analyse

Nettoyage du Data Frame

Remplace NaN values

Quartille Outliers

Analyse univariée, bivariée et multivariée

POC

1. IDEE D’APPLICATION

Input: Genre, poids, taille, âge.

Calcul du Basal Metabolic Rate (BMR); 3 possibilités:

- à partir des équation de Harris-Benedict;
- le laisser insérer à l’utilisateur;
- un mix des deux.

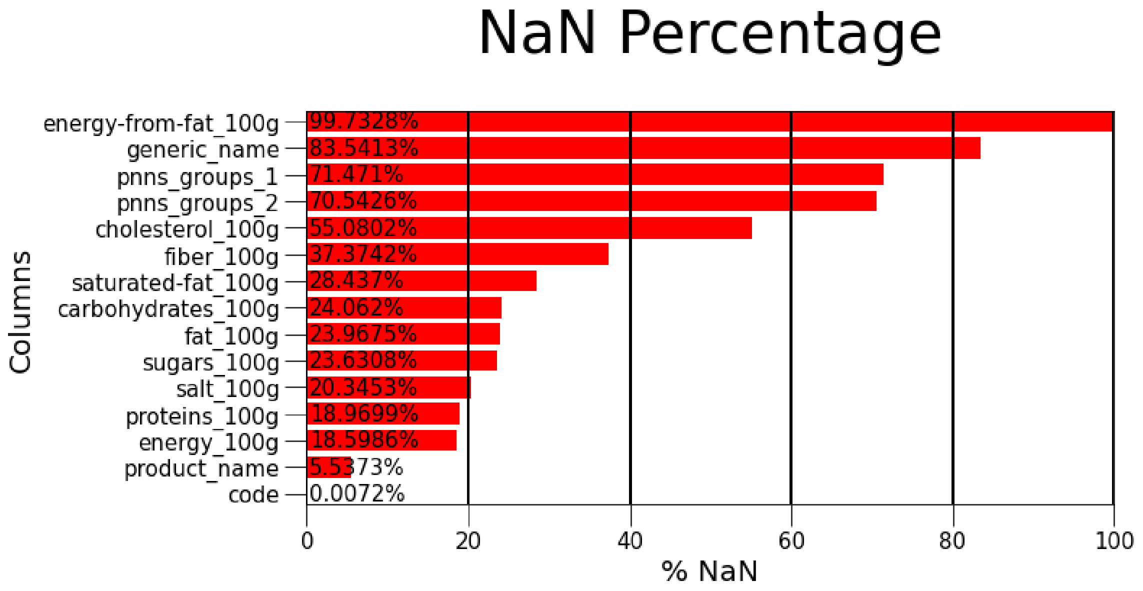
A chaque repas scanner les produits utilisés en insérant la quantité.

Pour faciliter cette tâche, créer son propre garde-manger et des convertion du style: une cuillère à soupe, une pincée etc.

L’application nous dit nos carences et nos surplus alimentaires.

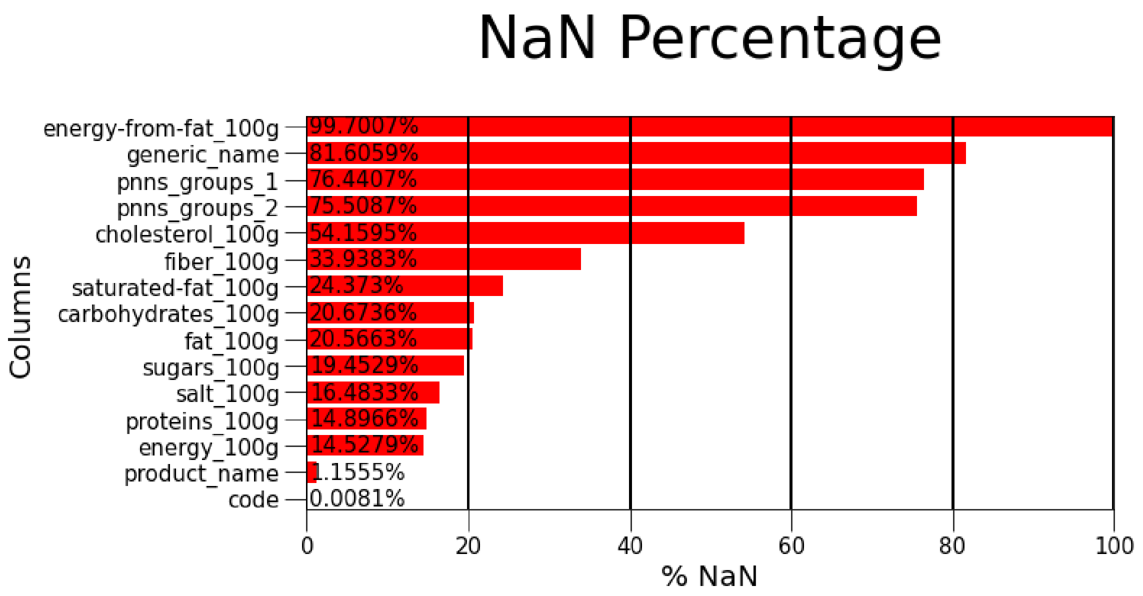
2. UNE PREMIÈRE ANALYSE

1. Le 76,22 % des valeurs sont nulles;
2. Je supprime le colonnes complètement vides ;
3. Après avoir vu la distribution des valeurs nulles par catégorie je choisie les features qui vont être utiles pour la création de mon application;



3. NETTOYAGE DU DATA FRAME

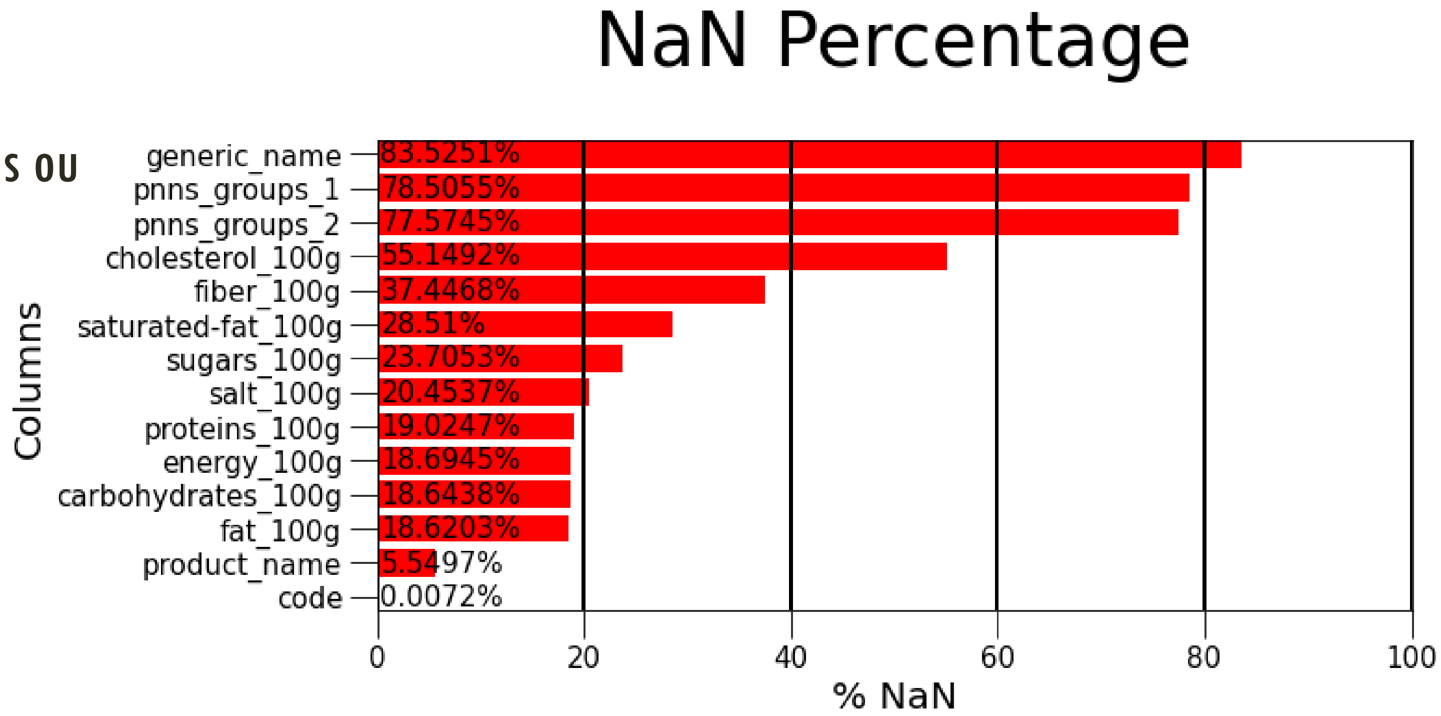
1. Drop energy-from-fat > energy;
2. Drop saturated-fat > fat;
3. Drop sugar > carbhoydrates;
4. Drop duplicated;
5. Remplace les valeurs aberrantes avec NaN;
6. Remplace la valeur ‘unknown’ avec NaN

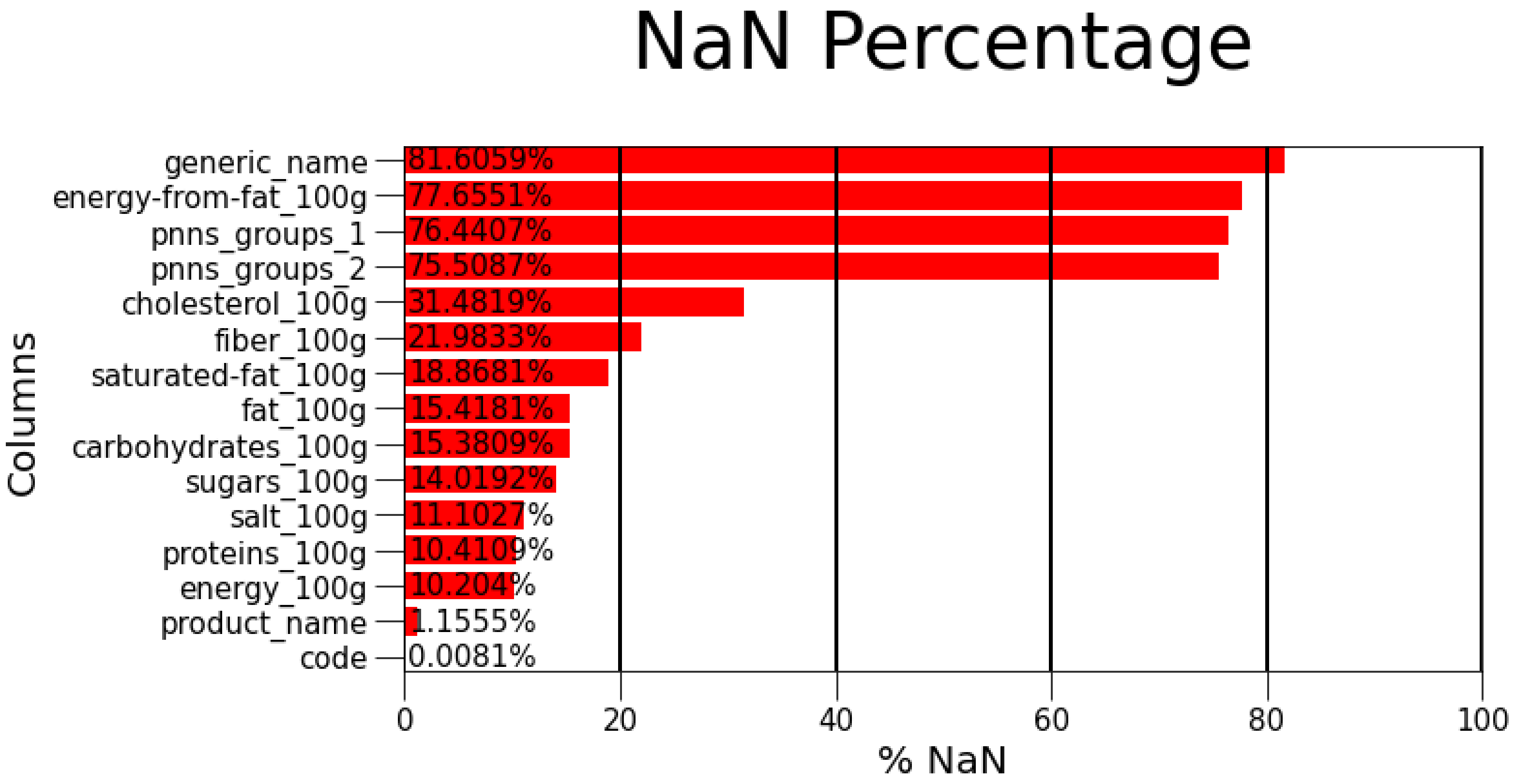


4. REMPLACE NAN VALUES

4.1 remplace avec des valeurs similaires ou par zéro

1. Energy from energy-from-fat;
2. Fat from saturated-fat;
3. Carbohydrates from sugar;
4. Si la somme des macronutriments est 100 je remplace les valeurs NaN avec zero.





4. REMPLACE NAN VALUES

4.2 remplace avec la moyenne de leur pnns

4. REMPLACE NAN VALUES

4.1 Sklearn methods

1. KNN Imputer
2. Iterative imputer
3. Simple Imputer

Data Frame	Size après drop_macronutrients_range
Data_knn_X_scaled	61847
Data_knn_X	61823
Data_ite_imp	62221
Data_simp_imp	61940
Data	60993

Méthode	Temp (min)
Knn_X_scaled	2416.07
Knn_X	2333.96
Ite_imp	0.1226
Simp_imp	0.0015

DESCRIBE MEAN

	energy_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
data_knn_X_scaled	8.966750e-01	1.293830e-02	3.944591e-01	1.065809e-03	1.239906e-02	3.399495e-01	1.866752e-01	8.441628e-03	-4.381985e-03
data_knn_X	7.959261e-01	1.028784e-02	3.958510e-01	1.452035e-03	1.166080e-02	3.580134e-01	1.854004e-01	7.199175e-03	2.198744e-04
data_ite_imp	-1.047140e+00	-3.568533e-03	-2.351983e-02	-1.666679e-03	-8.666548e-03	-1.398835e-01	-2.800383e-02	-7.305144e-03	-2.623944e-03
data_simp_imp	1.182116e-09	-7.972289e-12	5.547562e-12	-3.642225e-14	-4.168399e-11	-5.471179e-13	-2.920775e-12	5.240253e-13	2.251976e-12

	energy_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
data_knn_X_scaled	0.943428	1.166170	0.995745	0.705899	0.978295	0.930864	1.001468	1.008134	1.404554
data_knn_X	0.814163	0.780960	1.002581	1.025513	0.893830	1.014759	0.988784	0.811120	1.002308
data_ite_imp	1.550554	1.232885	1.056915	1.555320	1.431835	1.297645	1.134435	1.488811	0.485064
data_simp_imp	0.207037	0.714245	0.941411	0.176091	0.440290	0.647978	0.855817	0.330444	0.887310

Data Frame	SUM
data_knn_X_scaled	6.502049
data_knn_X	5.291166
data_ite_imp	7.323584
data_simp_imp	3.435142

DESCRIBE MEDIAN

	energy_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
data_knn_X_scaled	-39.465419	-2.3	-0.16	-0.001027	-6.25	-1.67	-0.2	-1.1	-0.16002
data_knn_X	-39.465419	-2.3	-0.17	-0.001982	-6.25	-1.67	-0.2	-1.1	-0.15818
data_ite_imp	-39.465419	-2.4	-1.62	-0.019712	-6.25	-3.17	-1.0	-1.1	-0.17818
data_simp_imp	-39.465419	-2.4	-1.62	-0.019712	-6.25	-3.17	-1.0	-1.1	-0.17818

	energy_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
data_knn_X_scaled	1.732051	1.0	1.006861	1.051748	0.0	1.0	1.0	0.0	0.901470
data_knn_X	0.577350	1.0	0.993115	0.946877	0.0	1.0	1.0	0.0	1.093896
data_ite_imp	0.577350	1.0	0.999988	0.999312	0.0	1.0	1.0	0.0	0.997683
data_simp_imp	0.577350	1.0	0.999988	0.999312	0.0	1.0	1.0	0.0	0.997683

Data Frame	SUM
data_knn_X_scaled	4.644521
data_knn_X	3.67124
data_ite_imp	3.575033
data_simp_imp	3.57033

DESCRIBE MEAN

	energy_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
data_knn_X_scaled_droped	3.734796	0.122546	0.380616	-0.001174	0.083882	0.519576	0.286354	0.003910	-0.016339
data_knn_X_droped	3.457381	0.104734	0.382274	-0.000397	0.099203	0.561245	0.277832	0.002706	-0.016514
data_ite_imp_droped	8.457727	0.317341	0.175119	-0.002558	-0.137213	-0.112187	0.151543	0.055764	-0.039327
data_simp_imp_droped	9.656012	0.270221	0.192557	-0.001273	-0.132403	0.183339	0.223041	0.064956	-0.040041

	energy_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
data_knn_X_scaled_droped	0.937384	0.884006	0.989667	0.227554	0.931141	0.846476	0.960182	0.972501	1.007253
data_knn_X_droped	1.037722	1.078006	1.006416	1.229964	1.066347	0.998785	0.801785	1.014428	0.992247
data_ite_imp_droped	0.770850	1.237610	1.086112	1.557381	1.019968	1.462729	1.545411	0.833414	0.969037
data_simp_imp_droped	1.204257	0.724402	0.909970	0.099863	0.977520	0.382532	0.216557	1.153515	1.030463

Data Frame	SUM
data_knn_X_scaled_droped	5.692468
data_knn_X_droped	5.990535
data_ite_imp_droped	6.376290
data_simp_imp_droped	5.306714

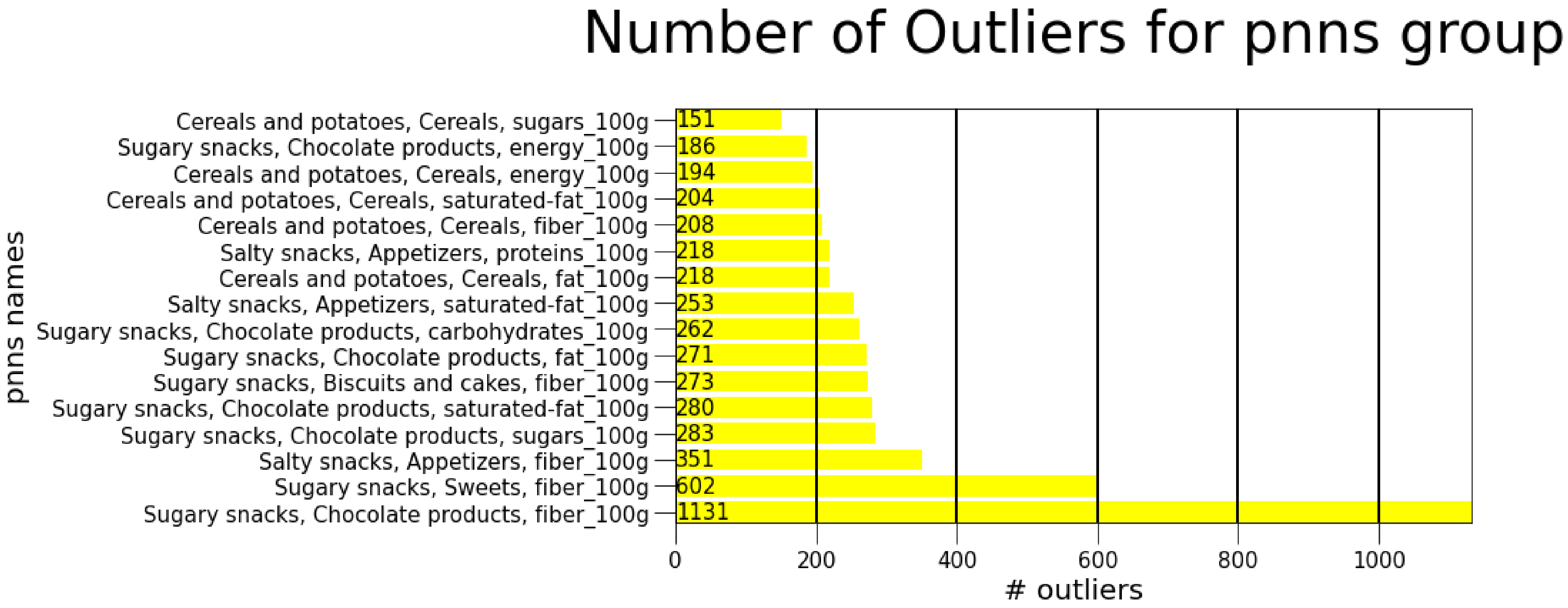
DESCRIBE MEDIAN

	energy_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
data_knn_X_scaled_droped	0.0	0.24	0.360000	0.0	0.05	0.81	0.0	0.0	0.00038
data_knn_X_droped	0.0	0.19	0.360000	0.0	0.05	0.89	0.0	0.0	-0.00254
data_ite_imp_droped	6.0	0.69	0.295845	0.0	-0.26	0.17	0.0	0.0	-0.00762
data_simp_imp_droped	6.0	0.69	0.295845	0.0	-0.26	1.81	0.0	0.0	-0.00762

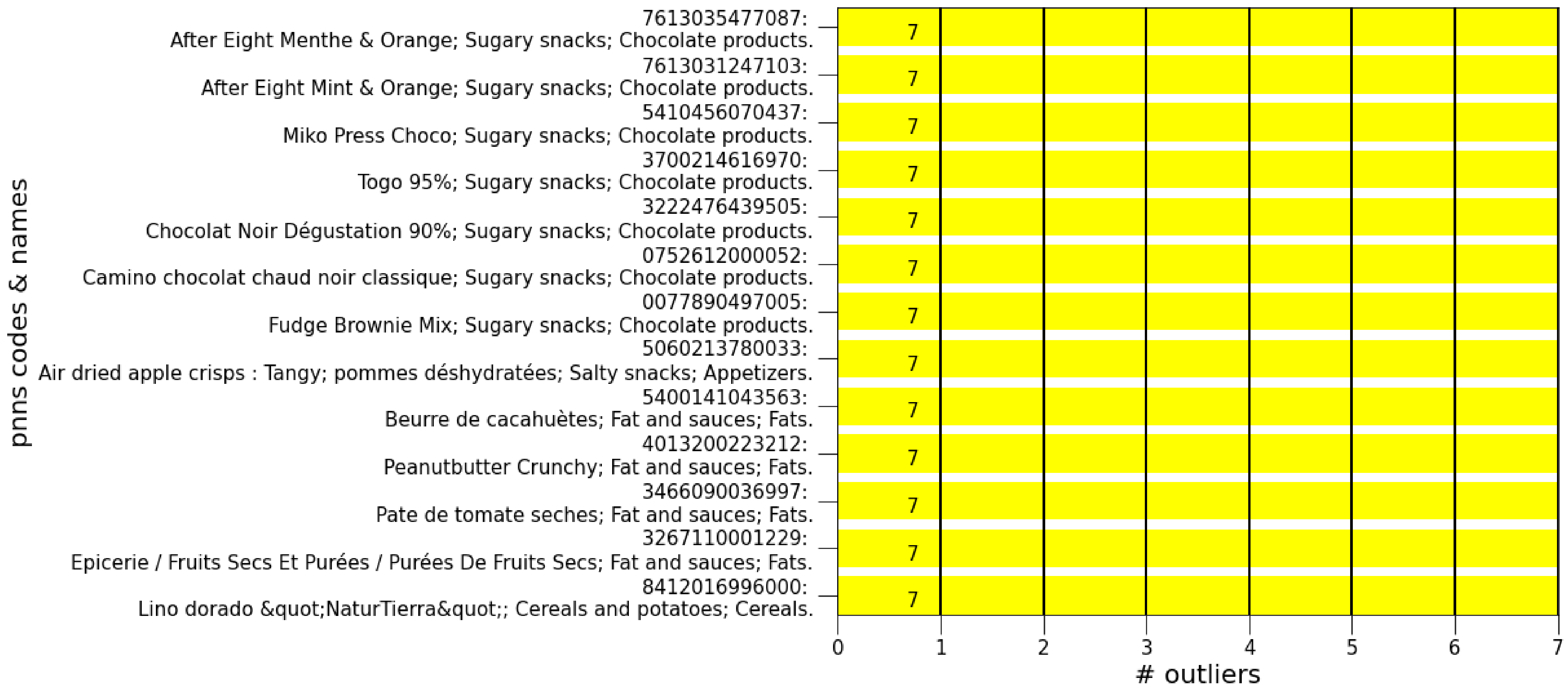
	energy_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
data_knn_X_scaled_droped	1.0	0.892269	1.0	0.0	1.0	0.188124	0.0	0.0	1.379372
data_knn_X_droped	1.0	1.102214	1.0	0.0	1.0	0.051306	0.0	0.0	0.527836
data_ite_imp_droped	1.0	0.997241	1.0	0.0	1.0	1.282660	0.0	0.0	0.953604
data_simp_imp_droped	1.0	0.997241	1.0	0.0	1.0	1.522090	0.0	0.0	0.953604

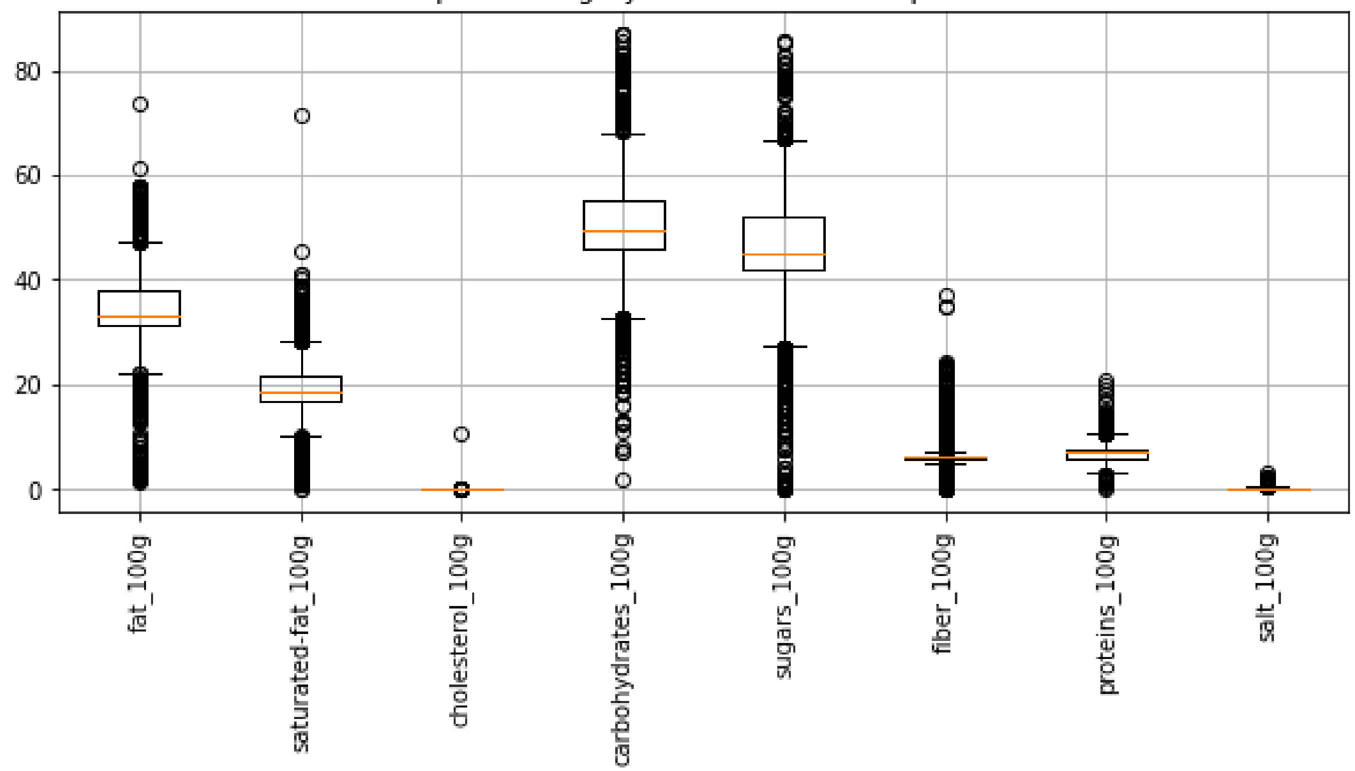
Data Frame	SUM
data_knn_X_scaled_droped	4.633521
data_knn_X_droped	3.671246
data_ite_imp_droped	3.575033
data_simp_imp_droped	3.575033

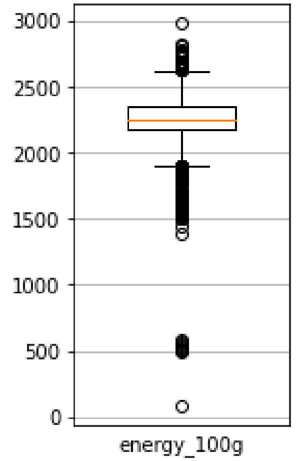
5. QUARTILLE OUTLIERS

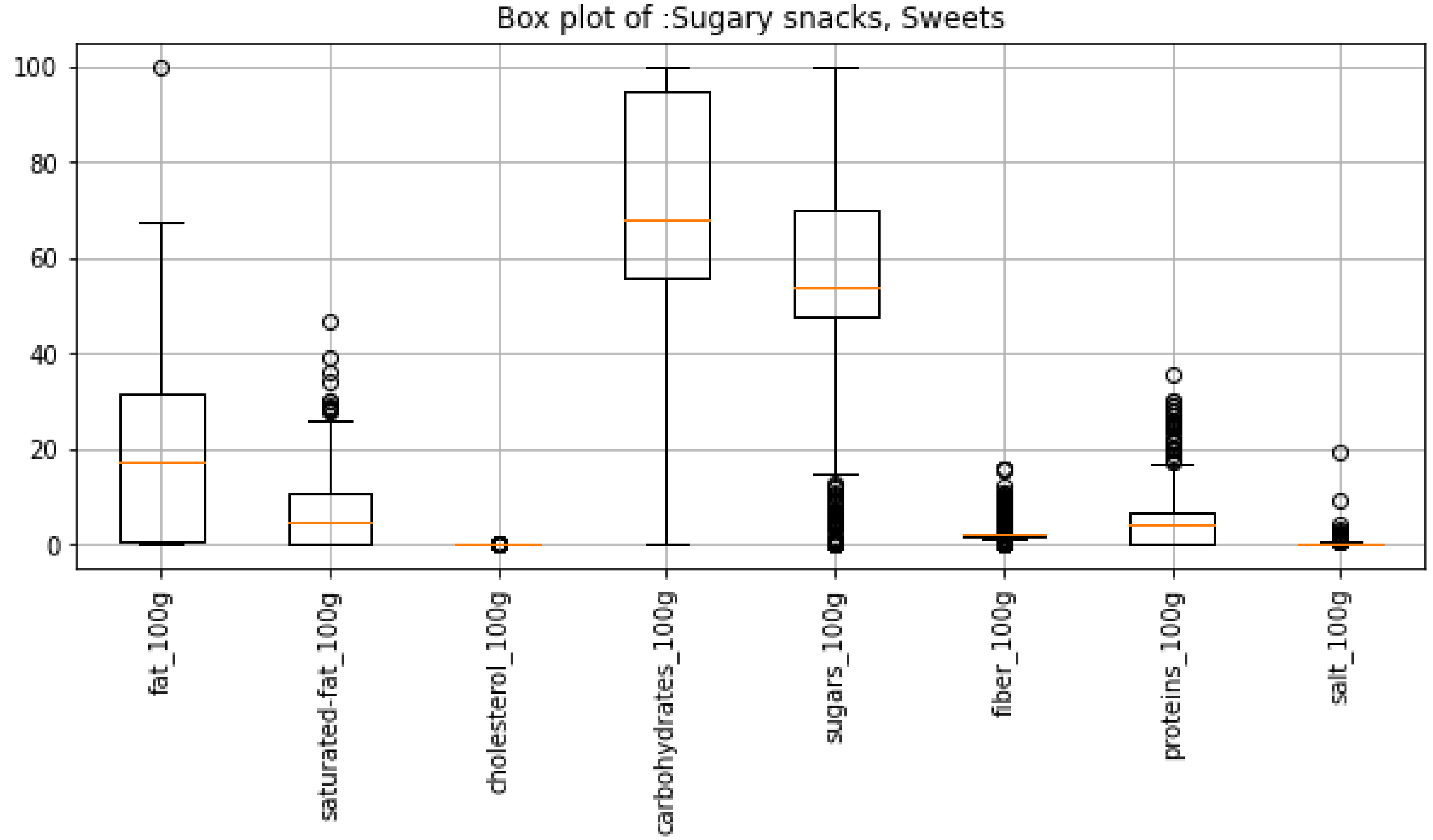


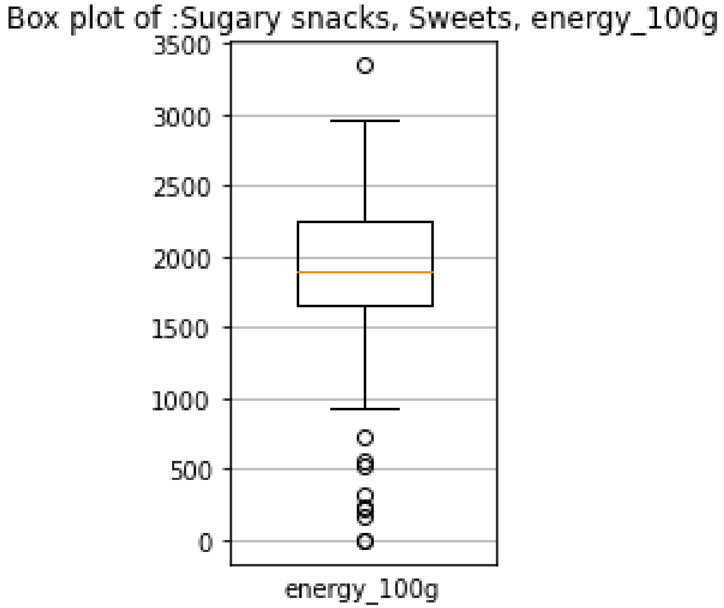
Products with more outliers elements



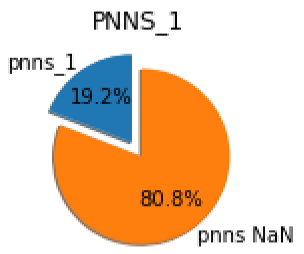

Box plot of :Sugary snacks, Chocolate products

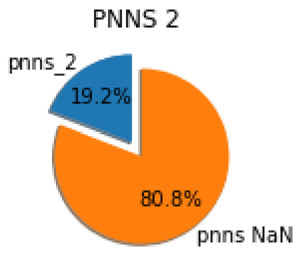

Box plot of :Sugary snacks, Chocolate products, energy_100g

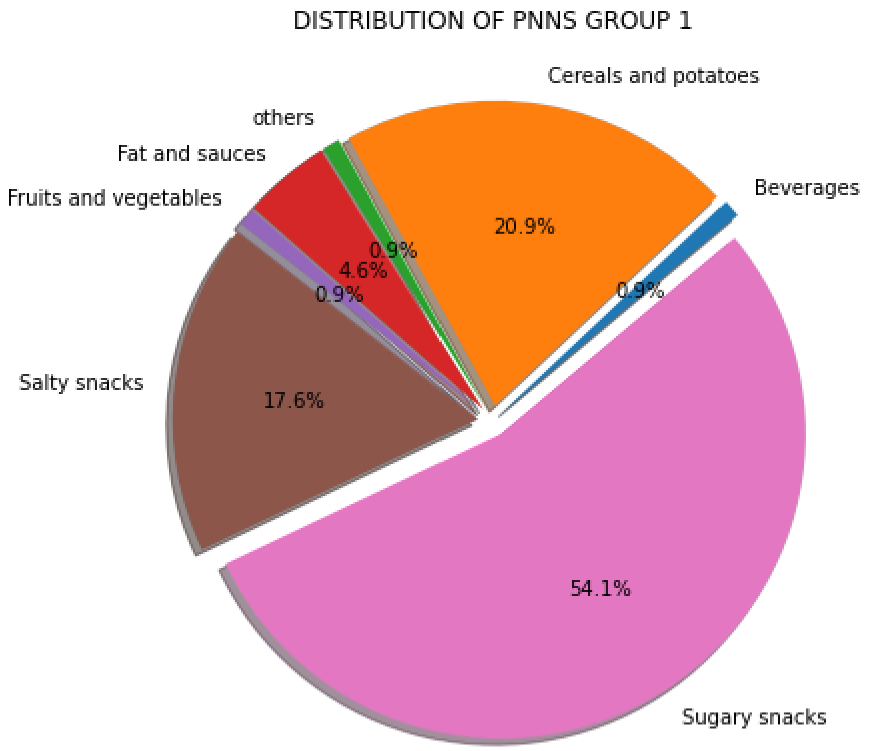

Box plot of :Sugary snacks, Sweets
100
80
60
40
20
0
fat_100g
saturated-fat_100g
cholesterol_100g
carbohydrates_100g
sugars_100g
fiber_100g
proteins_100g
salt_100g

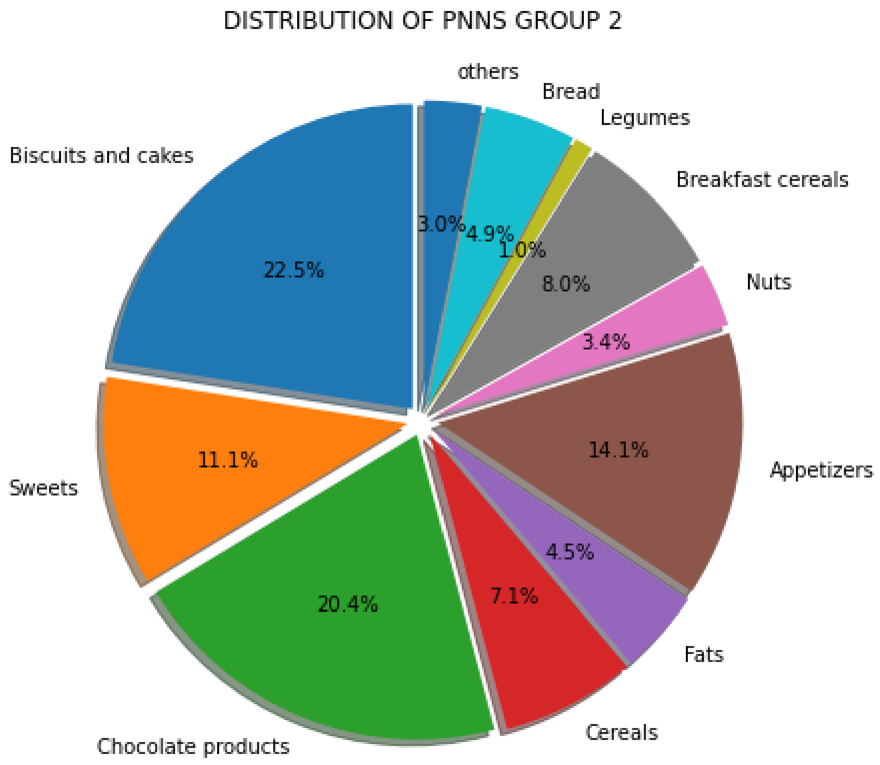

Box plot of :Sugary snacks, Sweets, energy_100g
3500
3000
2500
2000
1500
1000
500
0
energy_100g

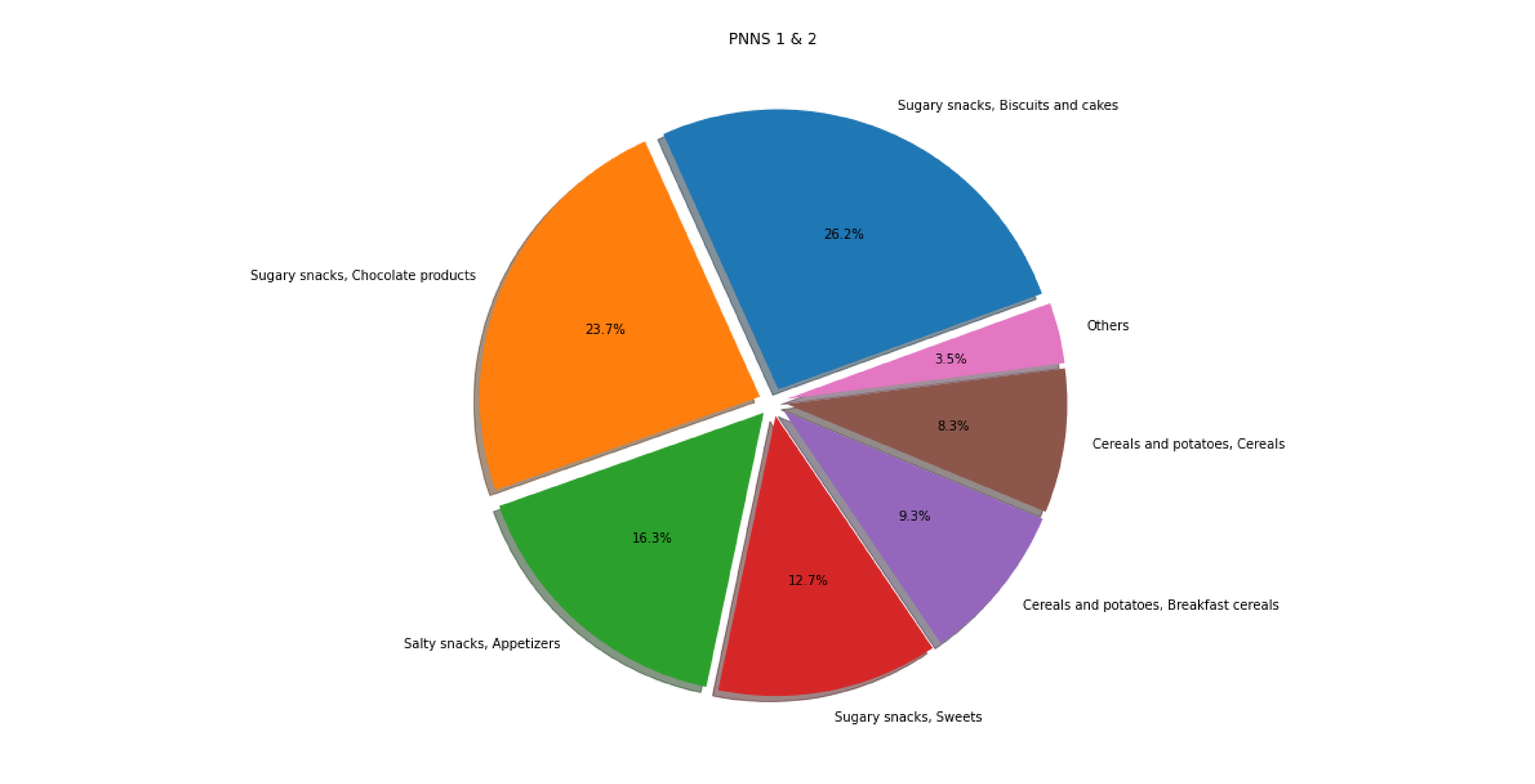
6. ANALYSE UNIVARIÉE ET BIVARIÉE


PNNS_1


PNNS 2

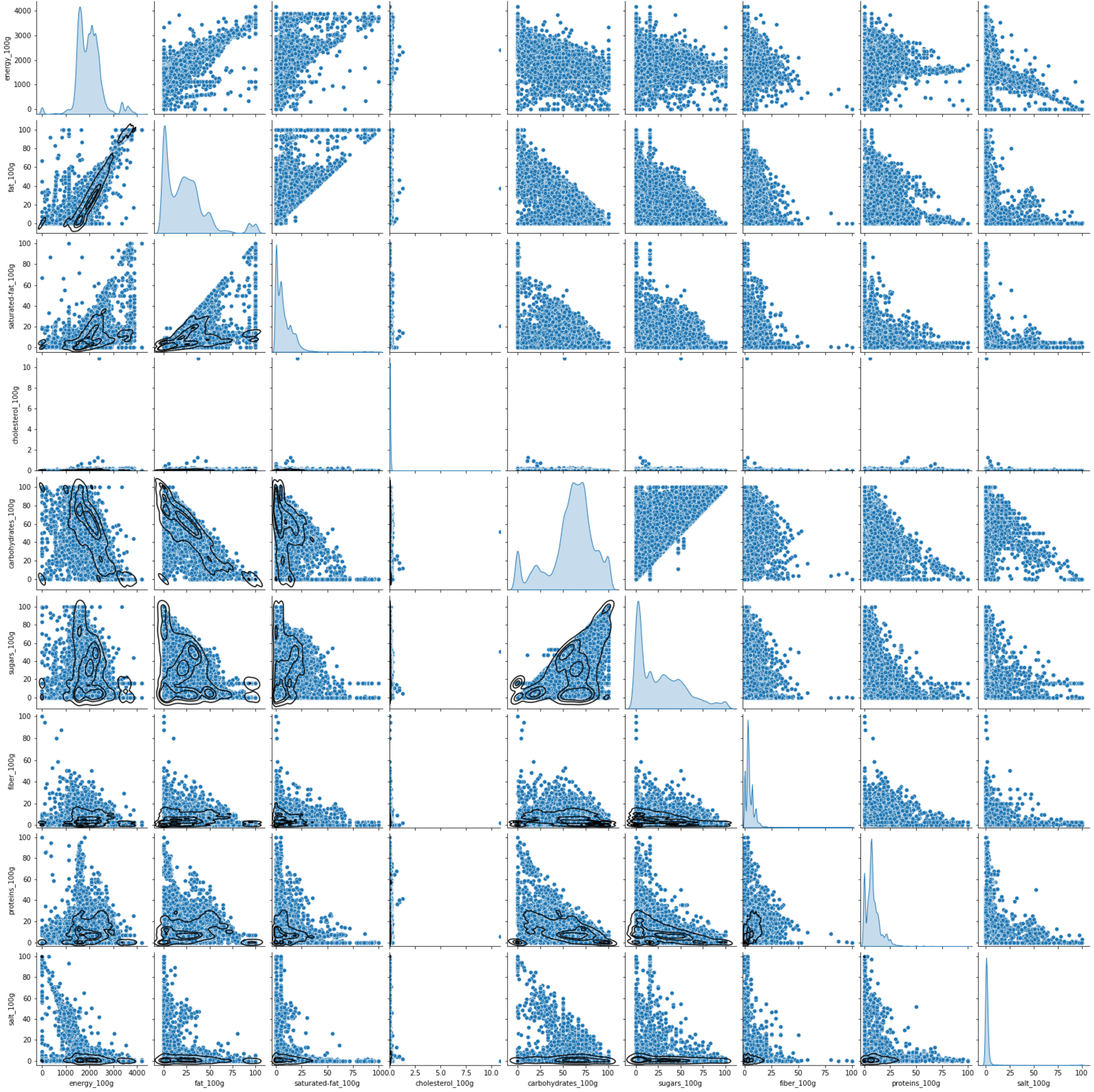

DISTRIBUTION OF PNNS GROUP 1


DISTRIBUTION OF PNNS GROUP 2

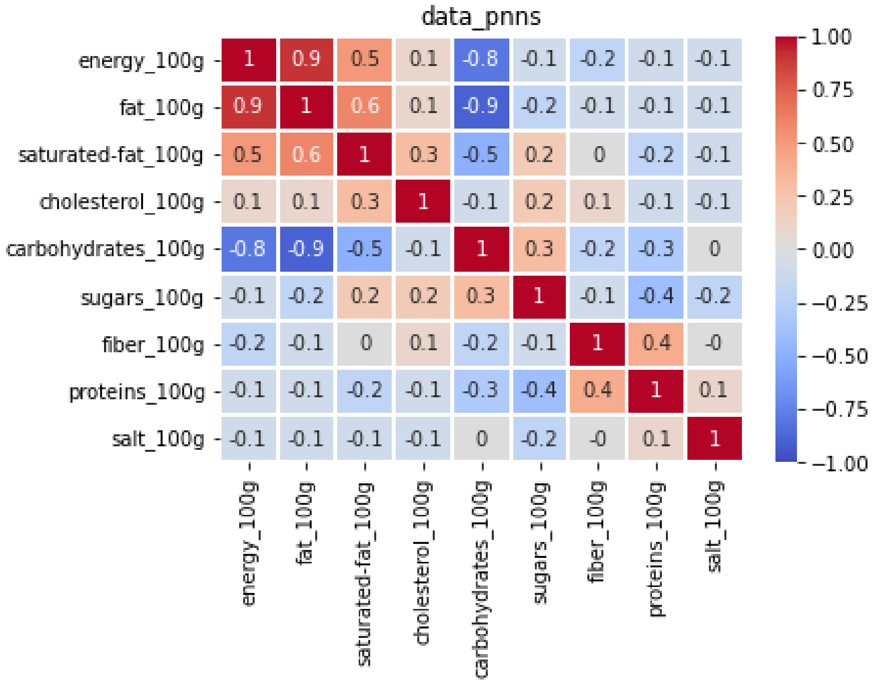

PNNS 1 & 2
Sugary snacks, Biscuits and cakes
26.2%
Sugary snacks, Chocolate products
23.7%
Others
3.5%
8.3%
Cereals and potatoes, Cereals
9.3%
16.3%
12.7%
Cereals and potatoes, Breakfast cereals
Salty snacks, Appetizers
Sugary snacks, Sweets

ANALYSE UNIVARIÉE ET BIVARIÉE

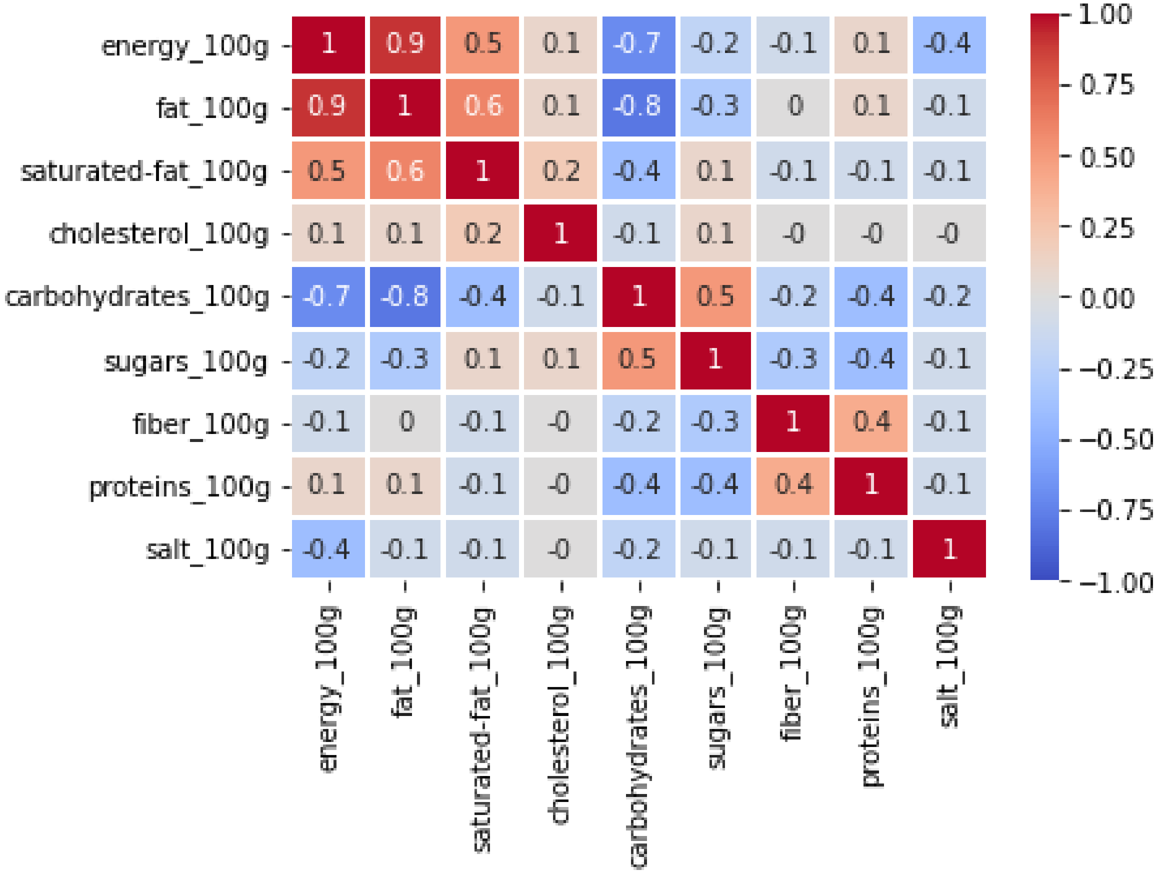
1. Energy_100g;
2. Fat_100g;
3. Saturated_fat_100g;
4. Cholesterol_100g;
5. Carbohydrates_100g;
6. Sugars_100g
7. Fiber_100g
8. Proteins_100g
9. Salt_100g



CORRELATIONS



	energy_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
energy_100g	1	0.9	0.5	0.1	-0.8	-0.1	-0.2	-0.1	-0.1
fat_100g	0.9	1	0.6	0.1	-0.9	-0.2	-0.1	-0.1	-0.1
saturated-fat_100g	0.5	0.6	1	0.3	-0.5	0.2	0	-0.2	-0.1
cholesterol_100g	0.1	0.1	0.3	1	-0.1	0.2	0.1	-0.1	-0.1
carbohydrates_100g	-0.8	-0.9	-0.5	-0.1	1	0.3	-0.2	-0.3	0
sugars_100g	-0.1	-0.2	0.2	0.2	0.3	1	-0.1	-0.4	-0.2
fiber_100g	-0.2	-0.1	0	0.1	-0.2	-0.1	1	0.4	-0
proteins_100g	-0.1	-0.1	-0.2	-0.1	-0.3	-0.4	0.4	1	0.1
salt_100g	-0.1	-0.1	-0.1	-0.1	0	-0.2	-0	0.1	1



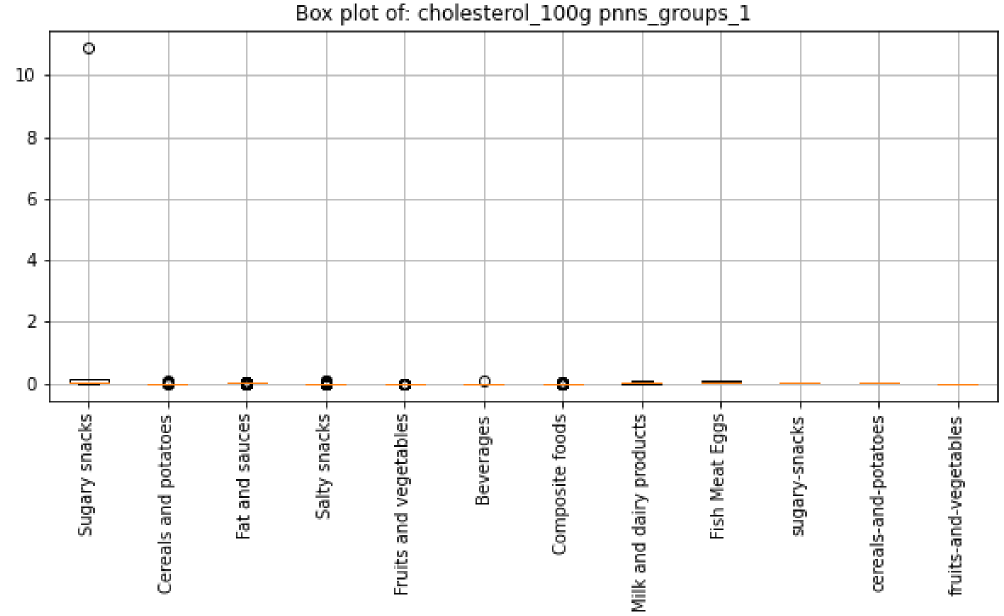
	energy_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
energy_100g	1	0.9	0.5	0.1	-0.7	-0.2	-0.1	0.1	-0.4
fat_100g	0.9	1	0.6	0.1	-0.8	-0.3	0	0.1	-0.1
saturated-fat_100g	0.5	0.6	1	0.2	-0.4	0.1	-0.1	-0.1	-0.1
cholesterol_100g	0.1	0.1	0.2	1	-0.1	0.1	-0	-0	-0
carbohydrates_100g	-0.7	-0.8	-0.4	-0.1	1	0.5	-0.2	-0.4	-0.2
sugars_100g	-0.2	-0.3	0.1	0.1	0.5	1	-0.3	-0.4	-0.1
fiber_100g	-0.1	0	-0.1	-0	-0.2	-0.3	1	0.4	-0.1
proteins_100g	0.1	0.1	-0.1	-0	-0.4	-0.4	0.4	1	-0.1
salt_100g	-0.4	-0.1	-0.1	-0	-0.2	-0.1	-0.1	-0.1	1

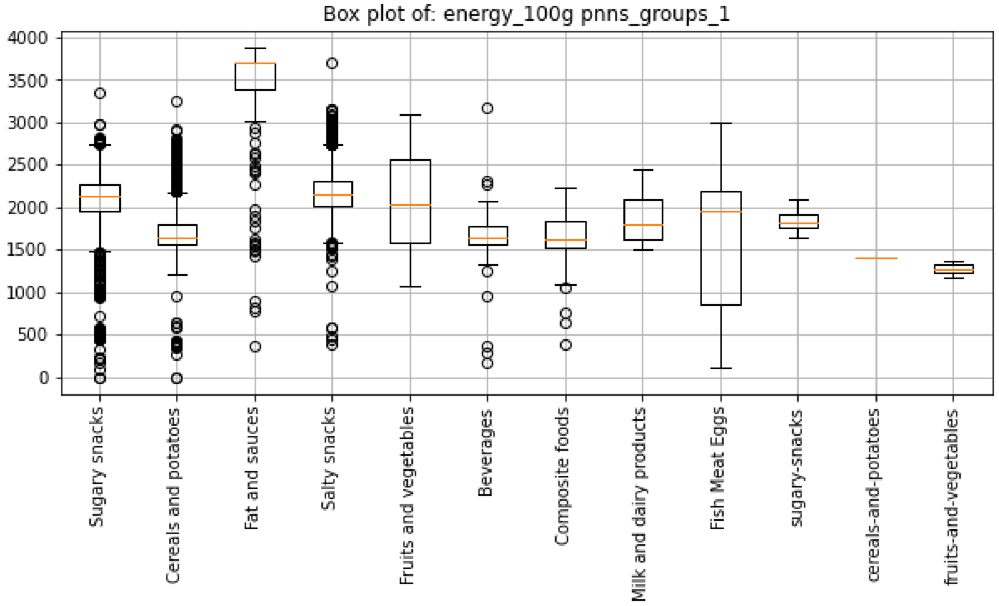
ANOVA PNNS 1

The ANOVA test has important assumptions that must be satisfied in order for the associated p-value to be valid:

1. The samples are independent.
2. Each sample is from a normally distributed population.
3. The population standard deviations of the groups are all equal. This property is known as homoscedasticity.

	F_statistics	p-values
cholesterol_100g	124.440153	4.509190e-272
energy_100g	1694.200128	0.000000e+00
fat_100g	2107.148301	0.000000e+00
saturated-fat_100g	491.663730	0.000000e+00
carbohydrates_100g	851.593744	0.000000e+00
sugars_100g	1573.172442	0.000000e+00
fiber_100g	200.518792	0.000000e+00
proteins_100g	403.426138	0.000000e+00
salt_100g	259.396382	0.000000e+00


Box plot of: cholesterol_100g pnns_groups_1
10
8
6
4
2
0
Sugary snacks
Cereals and potatoes
Fat and sauces
Salty snacks
Fruits and vegetables
Beverages
Composite foods
Milk and dairy products
Fish Meat Eggs
sugary-snacks
cereals-and-potatoes
fruits-and-vegetables

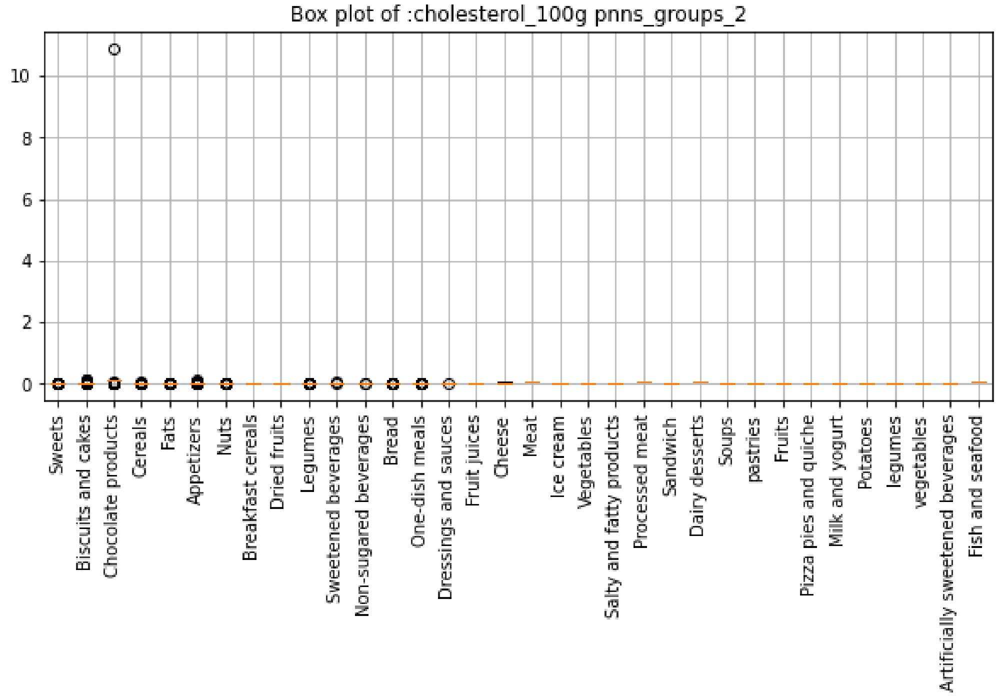

Box plot of: energy_100g pnns_groups_1
4000
3500
3000
2500
2000
1500
1000
500
0
Sugary snacks
Cereals and potatoes
Fat and sauces
Salty snacks
Fruits and vegetables
Beverages
Composite foods
Milk and dairy products
Fish Meat Eggs
sugary-snacks
cereals-and-potatoes
fruits-and-vegetables

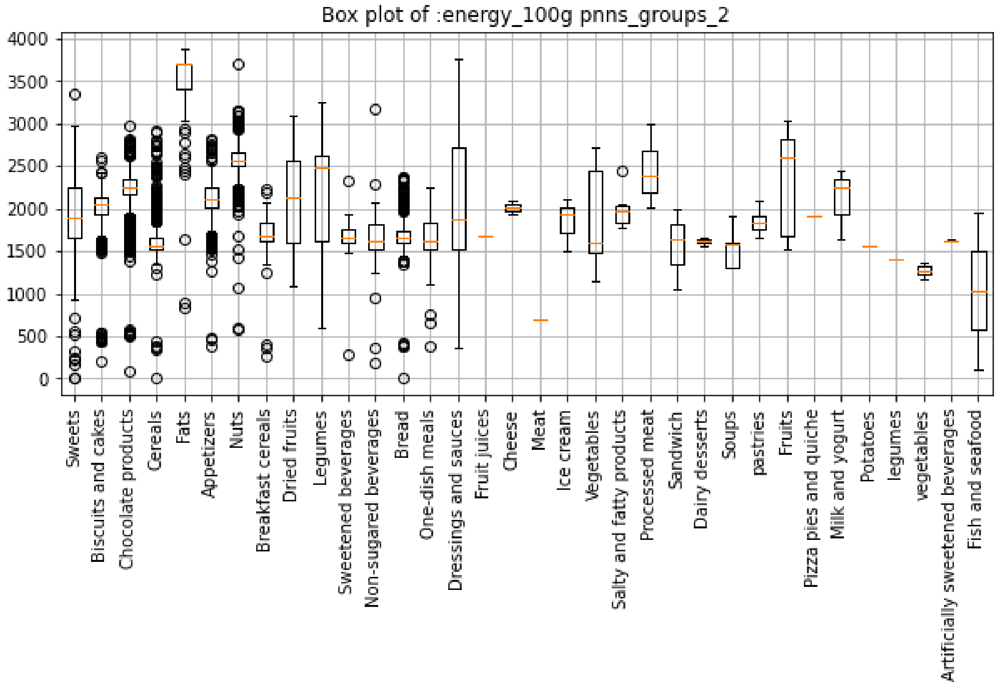
ANOVA PNNS 2

The ANOVA test has important assumptions that must be satisfied in order for the associated p-value to be valid:

1. The samples are independent.
2. Each sample is from a normally distributed population.
3. The population standard deviations of the groups are all equal. This property is known as homoscedasticity.

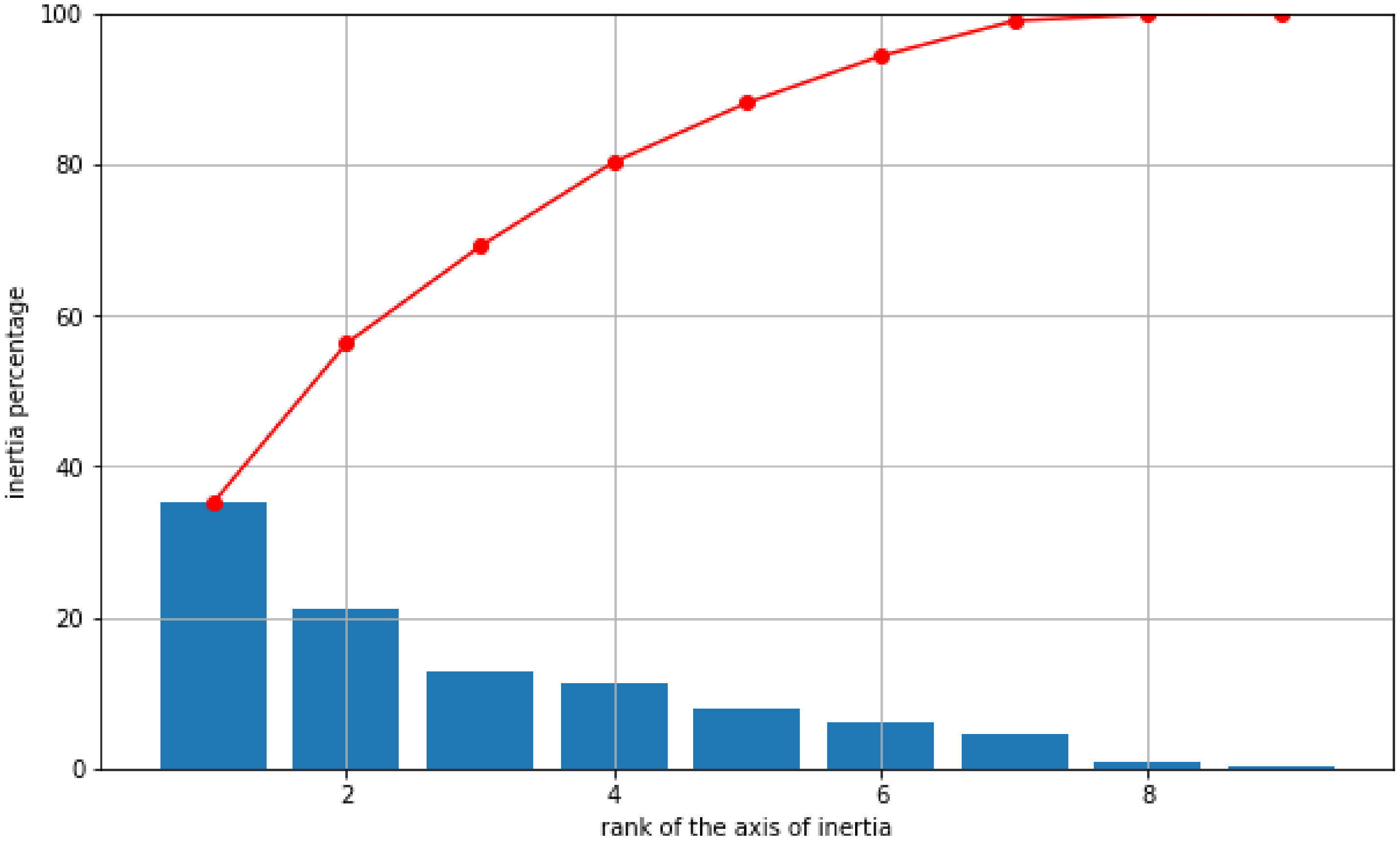
	F_statistics	p-values
energy_100g	880.076507	0.0
fat_100g	1307.931627	0.0
saturated-fat_100g	332.944747	0.0
cholesterol_100g	149.982729	0.0
carbohydrates_100g	774.537278	0.0
sugars_100g	957.154400	0.0
fiber_100g	193.778707	0.0
proteins_100g	462.492656	0.0
salt_100g	142.093494	0.0

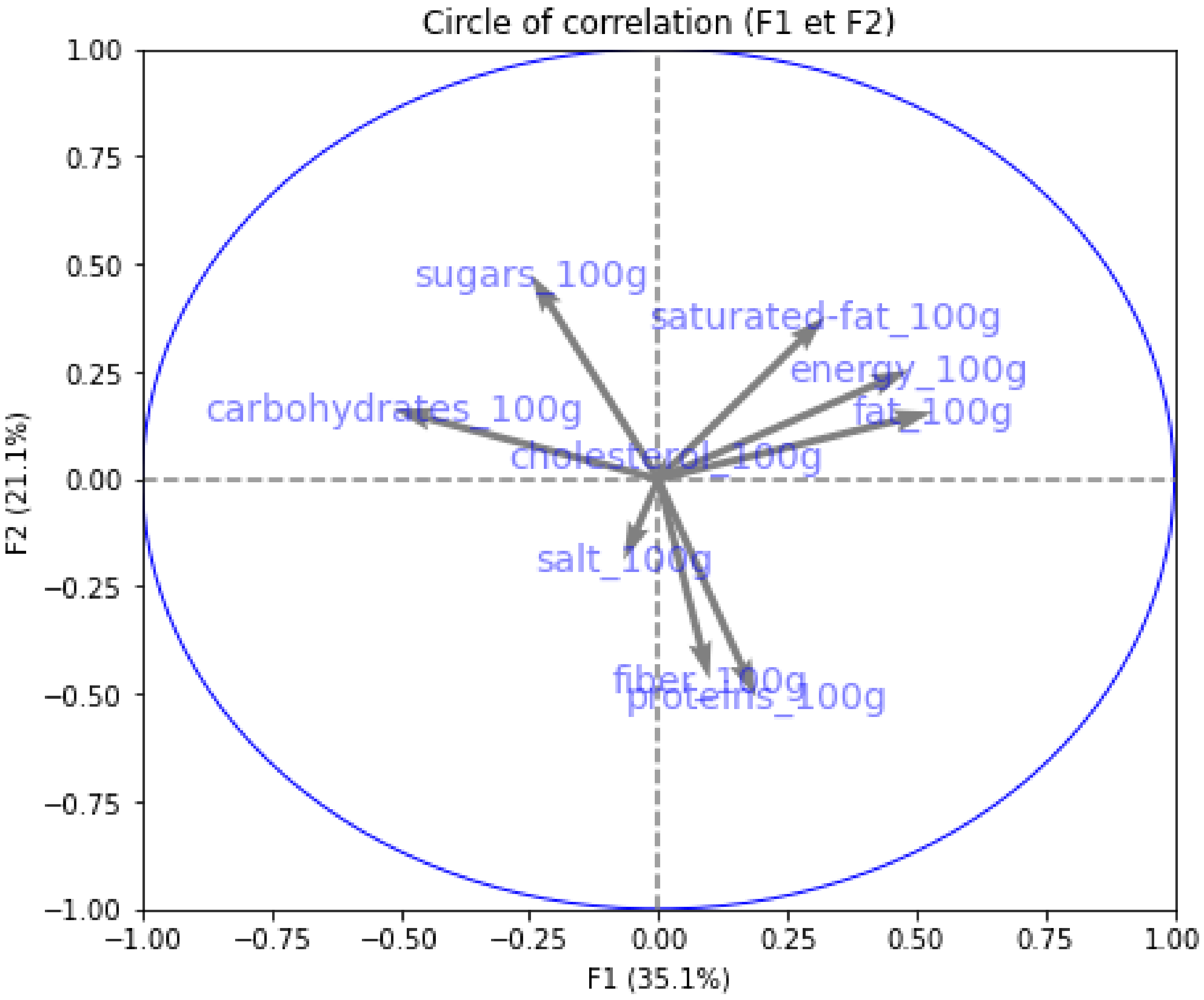

Box plot of :cholesterol_100g pnns_groups_2
0
2
4
6
8
10
Sweets
Biscuits and cakes
Chocolate products
Cereals
Fats
Appetizers
Nuts
Breakfast cereals
Dried fruits
Legumes
Sweetened beverages
Non-sugared beverages
Bread
One-dish meals
Dressings and sauces
Fruit juices
Cheese
Meat
Ice cream
Vegetables
Salty and fatty products
Processed meat
Sandwich
Dairy desserts
Soups
pastries
Fruits
Pizza pies and quiche
Milk and yogurt
Potatoes
legumes
vegetables
Artificially sweetened beverages
Fish and seafood


Box plot of :energy_100g pnns_groups_2
0
500
1000
1500
2000
2500
3000
3500
4000
Sweets
Biscuits and cakes
Chocolate products
Cereals
Fats
Appetizers
Nuts
Breakfast cereals
Dried fruits
Legumes
Sweetened beverages
Non-sugared beverages
Bread
One-dish meals
Dressings and sauces
Fruit juices
Cheese
Meat
Ice cream
Vegetables
Salty and fatty products
Processed meat
Sandwich
Dairy desserts
Soups
pastries
Fruits
Pizza pies and quiche
Milk and yogurt
Potatoes
legumes
vegetables
Artificially sweetened beverages
Fish and seafood

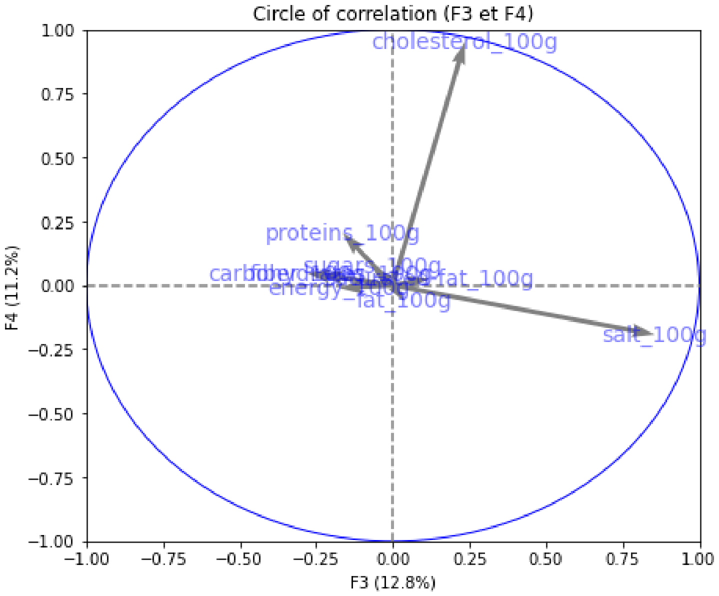
7. ANALYSE MULTIVARIÉE

PCA SCREE PLOT

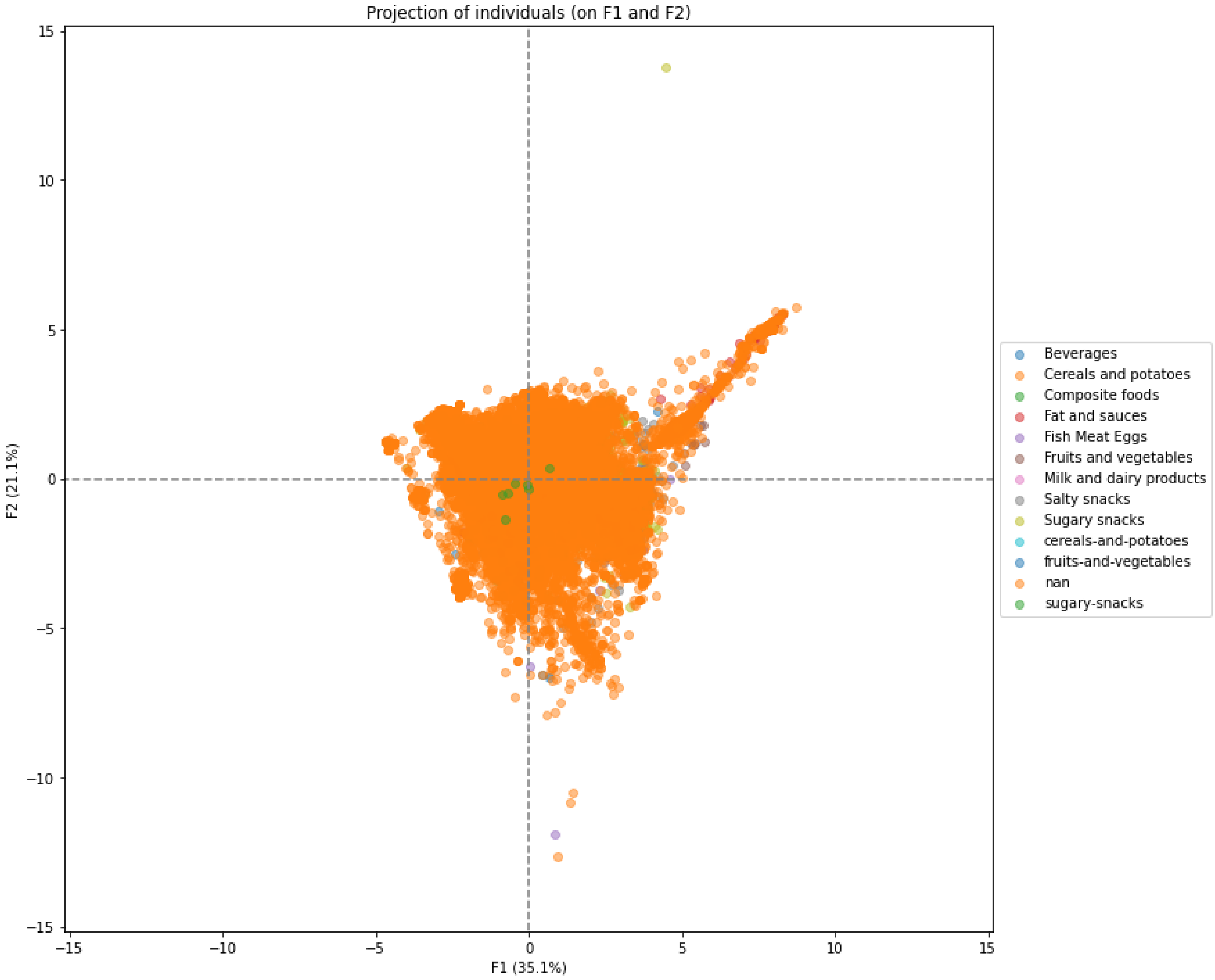




Circle of correlation (F1 et F2)



Circle of correlation (F3 et F4)


Projection of individuals (on F1 and F2)
F1 (35.1%)
F2 (21.1%)
−15
−10
−5
0
5
10
15
15
10
5
0
−5
−10
−15
Beverages
Cereals and potatoes
Composite foods
Fat and sauces
Fish Meat Eggs
Fruits and vegetables
Milk and dairy products
Salty snacks
Sugary snacks
cereals-and-potatoes
fruits-and-vegetables
nan
sugary-snacks

ANOVA MULTIVARIÉE

- Dataframe limité à la population munie de pnns;
- Souspopulation: répartition par quantitative_features

	F_statistics	p-values
Fish Meat Eggs, Fish and seafood	3.673917	1.040436e-02
fruits-and-vegetables, vegetables	154.296475	1.025263e-08
Composite foods, Sandwich	32.208205	3.958901e-09
Milk and dairy products, Cheese	527.446671	4.208567e-11
Milk and dairy products, Milk and yogurt	72.669018	4.027599e-12
Fish Meat Eggs, Processed meat	73.647053	3.587706e-12
Milk and dairy products, Ice cream	102.311970	2.052234e-13
Beverages, Artificially sweetened beverages	22671.398894	1.907964e-18
Milk and dairy products, Dairy desserts	3081.786483	6.548609e-38
Fruits and vegetables, Vegetables	161.245389	2.049360e-56
sugary-snacks, pastries	1121.670136	3.762518e-57
Fat and sauces, Dressings and sauces	96.026703	3.460496e-59
Fruits and vegetables, Soups	503.322143	5.592616e-60
Fruits and vegetables, Fruits	269.507003	1.268690e-82
Salty snacks, Salty and fatty products	1605.098789	1.585139e-108
Beverages, Non-sugared beverages	747.008490	5.655637e-251

8. POC

Une telle application n'est malheureusement pas réalisable avec ce jeu de donnée.

Une idée pour subvenir aux manques du Data Frame est celle de demander aux utilisateurs d'insérer les valeurs manquantes et de valider si le valeurs ajoutés pendant cette analyse sont correctes.