Analyse du jeu de données FoodFacts

Melanie Cosmides - Remy Zirnheld - Yohan Pipereau - Stevan Coroller

Le jeu de données Food Facts

Le jeu de données et l’objectif métier

Jeu de donnée: FoodFacts (https://world.openfoodfacts.org/data)

Ce que l’on recherche:
Si quelqu’un cherche un aliment avec des caractéristiques (composition nutritionnelle) particulière, peut-on toujours satisfaire sa demande ?

=> si c’est le cas, peut-on survivre uniquement en mangeant certains aliments ?

Peut-on créer des groupes d’aliments de composition proches, afin de construire des repas équilibrés ?

Organisation des données

On voit :

- des données diverses (ingrédients, creator, last_modified…)
- des données numériques (quantité de protéines pour 100g…)

Potentiel d'utilisation :

- Déterminer si on peut faire des classifications des aliments
- Prédire une caractéristique pour un aliment ? (pour compléter des données lacunaires par exemple)

Organisation des données

code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name
0000000000017	http://world-en.openfoodfacts.org/product/0000000000017/vitoria-crackers	kiliweb	1529059080	2018-06-15T10:38:00Z	1529059204	2018-06-15T10:40:04Z	Vitória crackers
0000000000031	http://world-en.openfoodfacts.org/product/0000000000031/cacao	isagoofy	1539464774	2018-10-13T21:06:14Z	1539464817	2018-10-13T21:06:57Z	Cacao
0000000000123	http://world-en.openfoodfacts.org/product/0000000000123/sauce-sweety-chili-0	kiliweb	1535737982	2018-08-31T17:53:02Z	1535737986	2018-08-31T17:53:06Z	Sauce Sweety chil
0000000000178	http://world-en.openfoodfacts.org/product/0000000000178/mini-coco	kiliweb	1542456332	2018-11-17T12:05:32Z	1542456333	2018-11-17T12:05:33Z	Mini coco
0000000000208	http://world-en.openfoodfacts.org/product/0000000000208/pistou-d-ail-des-ours	kiliweb	1544207680	2018-12-07T18:34:40Z	1544207683	2018-12-07T18:34:43Z	Pistou d'ail des
0000000000284	http://world-en.openfoodfacts.org/product/0000000000284/pain-mais	kiliweb	1547120245	2019-01-10T11:37:25Z	1547120246	2019-01-10T11:37:26Z	Pain maïs
0000000000291	http://world-en.openfoodfacts.org/product/0000000000291/mendiants	kiliweb	1534239669	2018-08-14T09:41:09Z	1534239732	2018-08-14T09:42:12Z	Mendiants
0000000000949	http://world-en.openfoodfacts.org/product/0000000000949/salade-de-carottes-rapees	kiliweb	1523440813	2018-04-11T10:00:13Z	1546194697	2018-12-30T18:31:37Z	Salade de carotte
0000000000970	http://world-en.openfoodfacts.org/product/0000000000970/fromage-blanc-aux-myrtilles	kiliweb	1520506368	2018-03-08T10:52:48Z	1520506371	2018-03-08T10:52:51Z	Fromage blanc aux
0000000001001	http://world-en.openfoodfacts.org/product/0000000001001	openfoodfacts-contributors	1537766416	2018-09-24T05:20:16Z	1537766416	2018-09-24T05:20:16Z	
0000000001007	http://world-en.openfoodfacts.org/product/0000000001007/vainilla	kiliweb	1538127563	2018-09-28T09:39:23Z	1538127565	2018-09-28T09:39:25Z	Vainilla
0000000001137	http://world-en.openfoodfacts.org/product/0000000001137/baguette-parisien	kiliweb	1539781575	2018-10-17T13:06:15Z	1539781578	2018-10-17T13:06:18Z	Baguette parisien

carbonate_100g	potassium_100g	chloride_100g	calcium_100g	phosphorus_100g	iron_100g	magnesium_100g	zinc_100g	copper_100g	manganese_100g	fluoride_100g	selenium_100g	chromium_100g	molybdenum_100g	iodine_100g	caffeine_100g	taurine_100g
		70.1	15										7.8			
		4.8	0.4										0.2			
		10	3										2			
								5.3	3.9							
		16.3	16.3										4.4			
		38.4	1.8										11.7			
				41	2										12.5	
												27.3	21.9			
		24	23										0			
		39	20										7.8			
		7.6	7.5										4.6			
						0	0									
		9.2	0.9										8.3			
		5.2	1.2										7			
		3.9	1										1.9			
		20.7	3.8										9.1			

Data cleaning

Le fichier CSV n’utilise pas un séparateur uniforme (whitespace, tabspace, double tabspace).

Les 66 première colonnes sont de type string.

Les temps de chargement de tout le fichier avec python prenne une vingtaine de secondes !

Construire le jeu de données à traiter

On remarque que:

- on n’a pas le même nombre de colonnes pour chaque produit
- on a beaucoup de cellules non remplies

On décide:

- de lire uniquement les colonnes qui nous intéressent
- de faire une analyse afin de déterminer sur quelles caractéristiques des produits on peut travailler (ACP)

Rappels sur l’ACP

Rappels sur l’ACP

- Soit un jeu de données à 3 colonnes (C1, C2, C3). Représentation triviale en 2D par nuages de point colorés

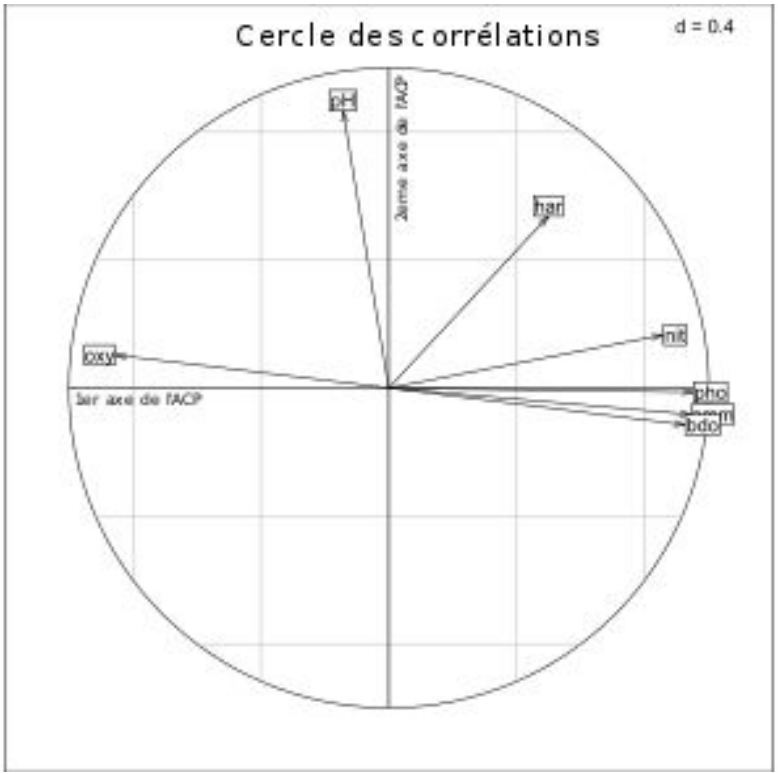
- Soit n colonnes (C1, …, Cn), comment représenter le jeu de données afin d’ évaluer au mieux la présence d’informations importantes ?

Résultat de l’ACP

- L’ACP permet de réaliser une réduction du problème d’un espace de dimension n en un espace de dimension 2 avec deux axes:
  - F1: axe principale
  - F2: axe secondaire

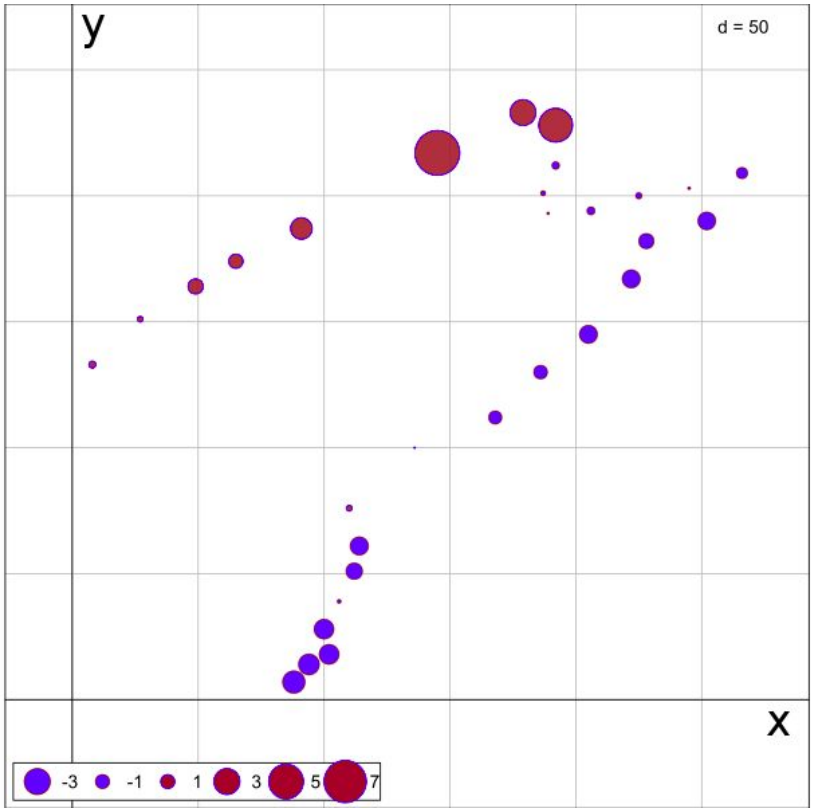
- On visualise les données sous la forme de 2 graphes
  - Le cercle des corrélations
  - Le graphe des observations

Résultats de l’ACP - Cercle des corrélations



- Contribution des colonnes Ci aux axes F1 et F2 générés
- Détermine si deux colonnes Ci et Cj:
  - corrélés positivement
  - corrélés négativement
  - indépendantes

Résultats de l’ACP - Graphique des observations

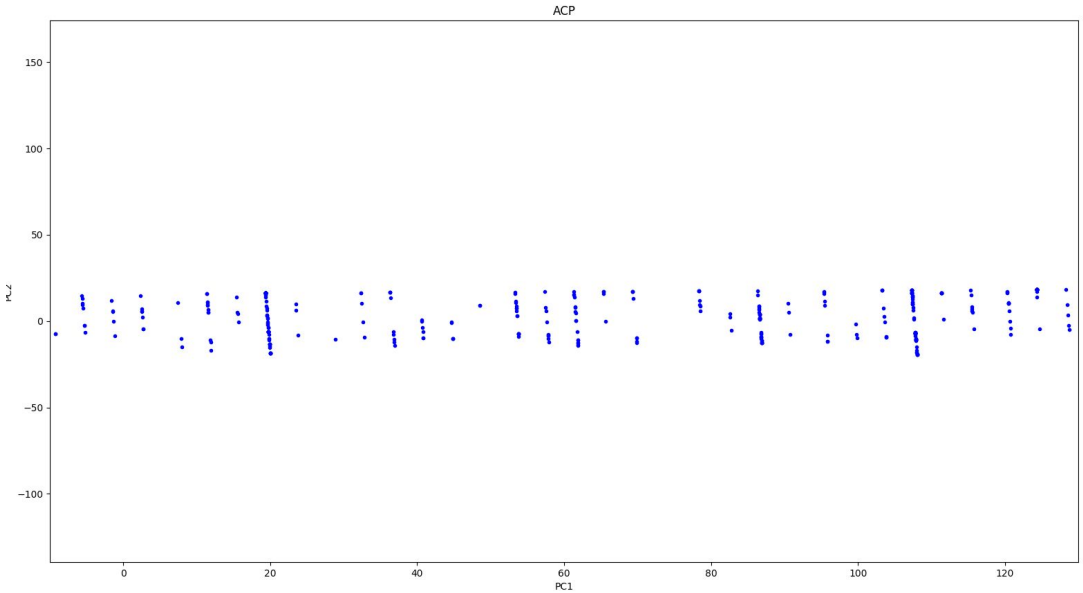


- Interprétation des données types sur l’axe (F1,F2) d’après le cercle des corrélations

Résultats ACP sur Food Facts

ACP sur energy - fat - saturated fat

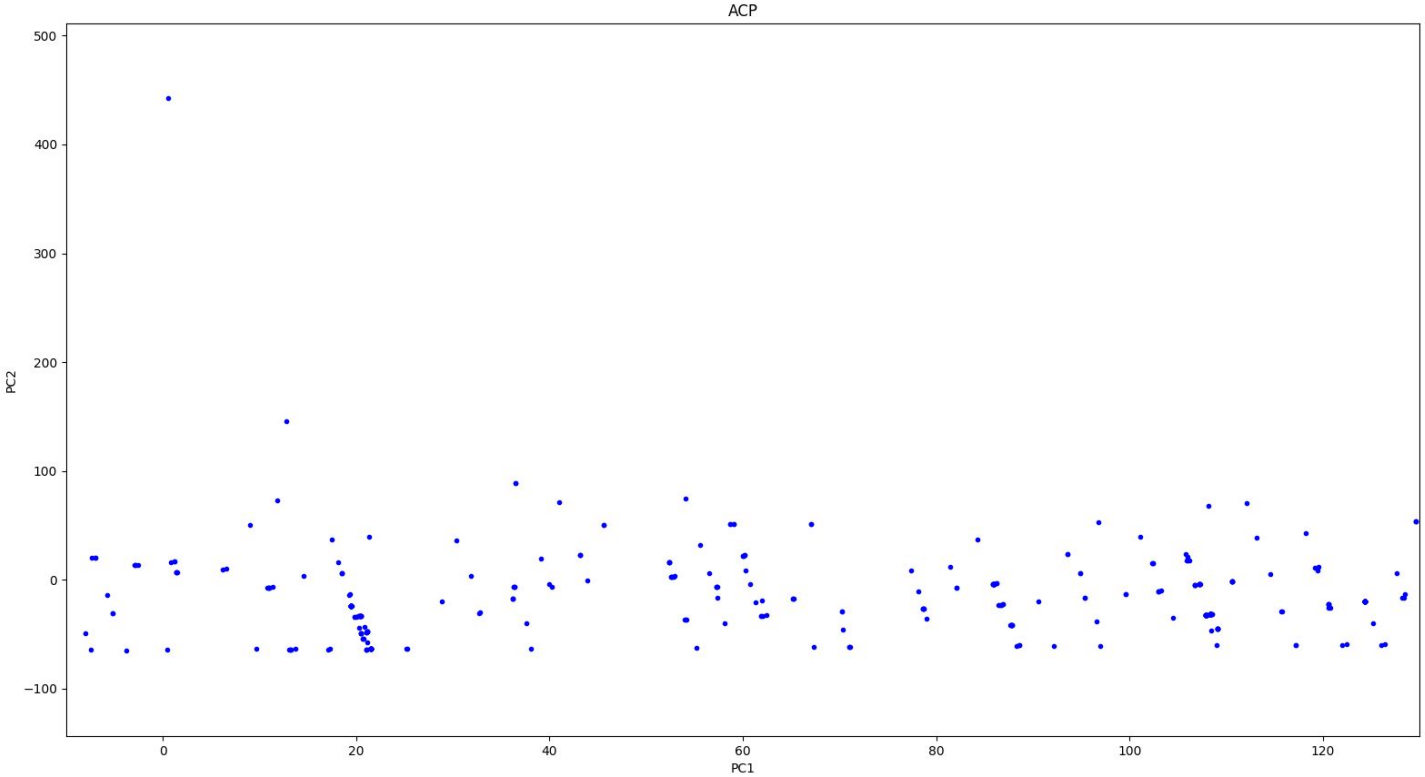
```
pca = PCA(n_components=2)
principal_components = pca.fit_transform(cleared_dataset)
principal_df = pd.DataFrame(data=principal_components, columns=['PC1', 'PC2’])
```



Pourcentage de contribution de chaque axe : 99.97%, 0.02%

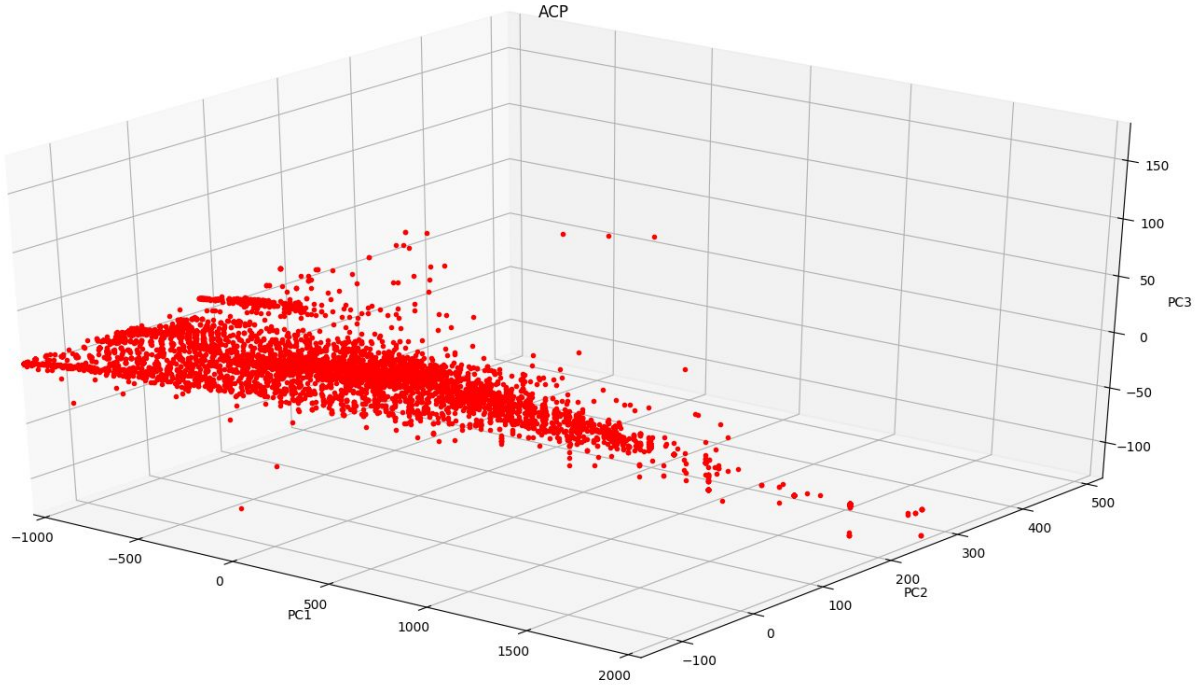
ACP

Pourcentage de contribution de chaque axe : 99.5%, 0.45%
['serving_quantity', 'energy_100g', 'fat_100g', 'saturated-fat_100g']



ACP

Pourcentage de contribution de chaque axe : 99.5%, 0.45%, 0.02%
['serving_quantity', 'energy_100g', 'fat_100g', 'saturated-fat_100g']



Rappels Kmeans

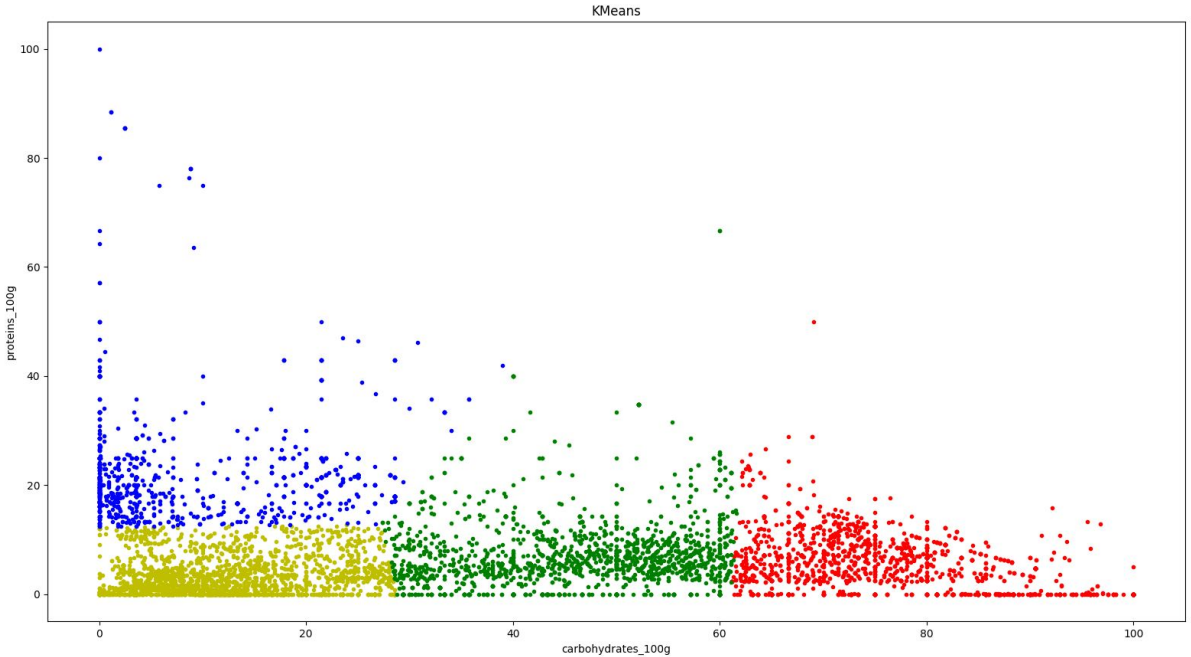
Rappels sur le KMeans

- on a des données représentées par x caractéristiques
- on initialise les centroïdes (centres des groupes qu’on souhaite trouver)
- par itérations successives on cherche à déterminer des groupes de données

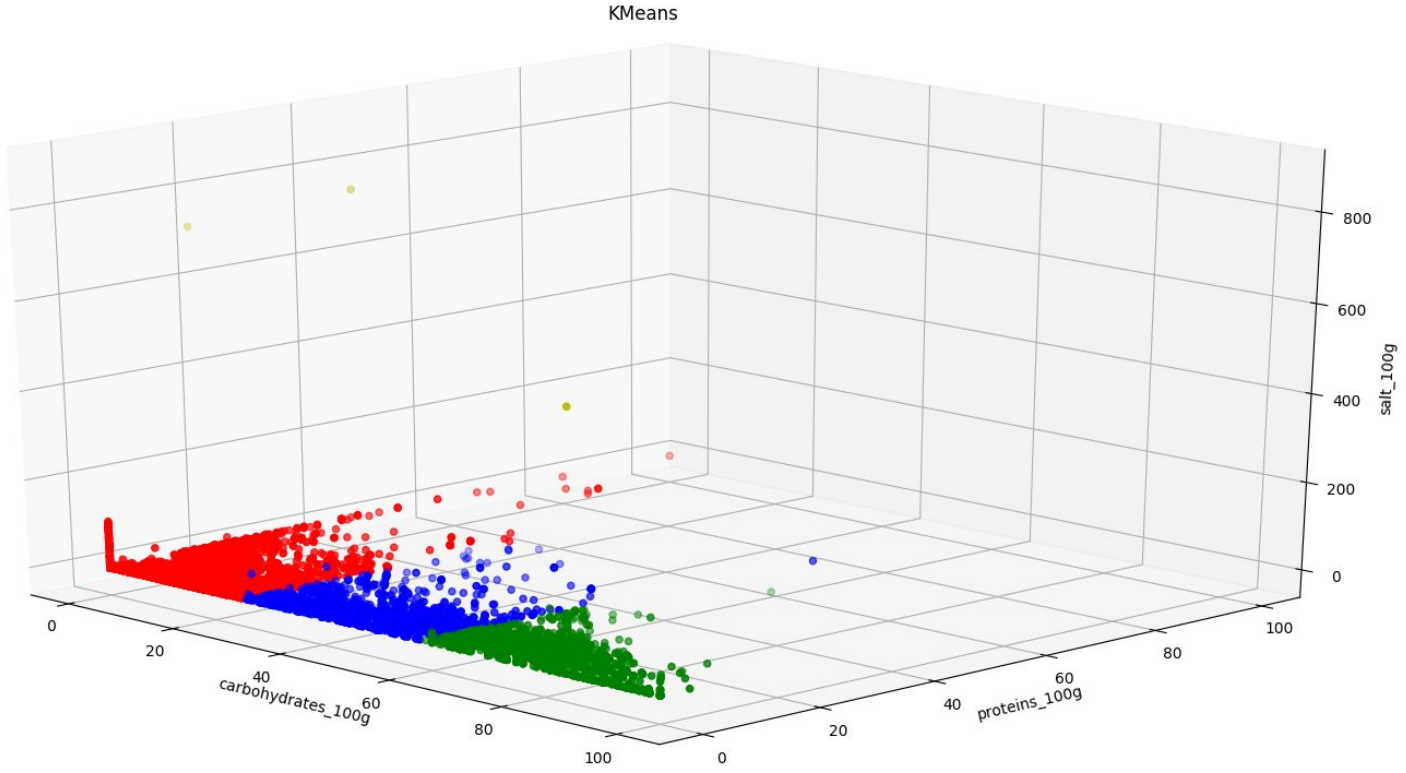
Résultats Kmeans sur Food Facts

KMeans

```
kmeans = KMeans(n_clusters=nb_clusters, random_state=0).fit(cleared_dataset)
pred = kmeans.predict(data)
```



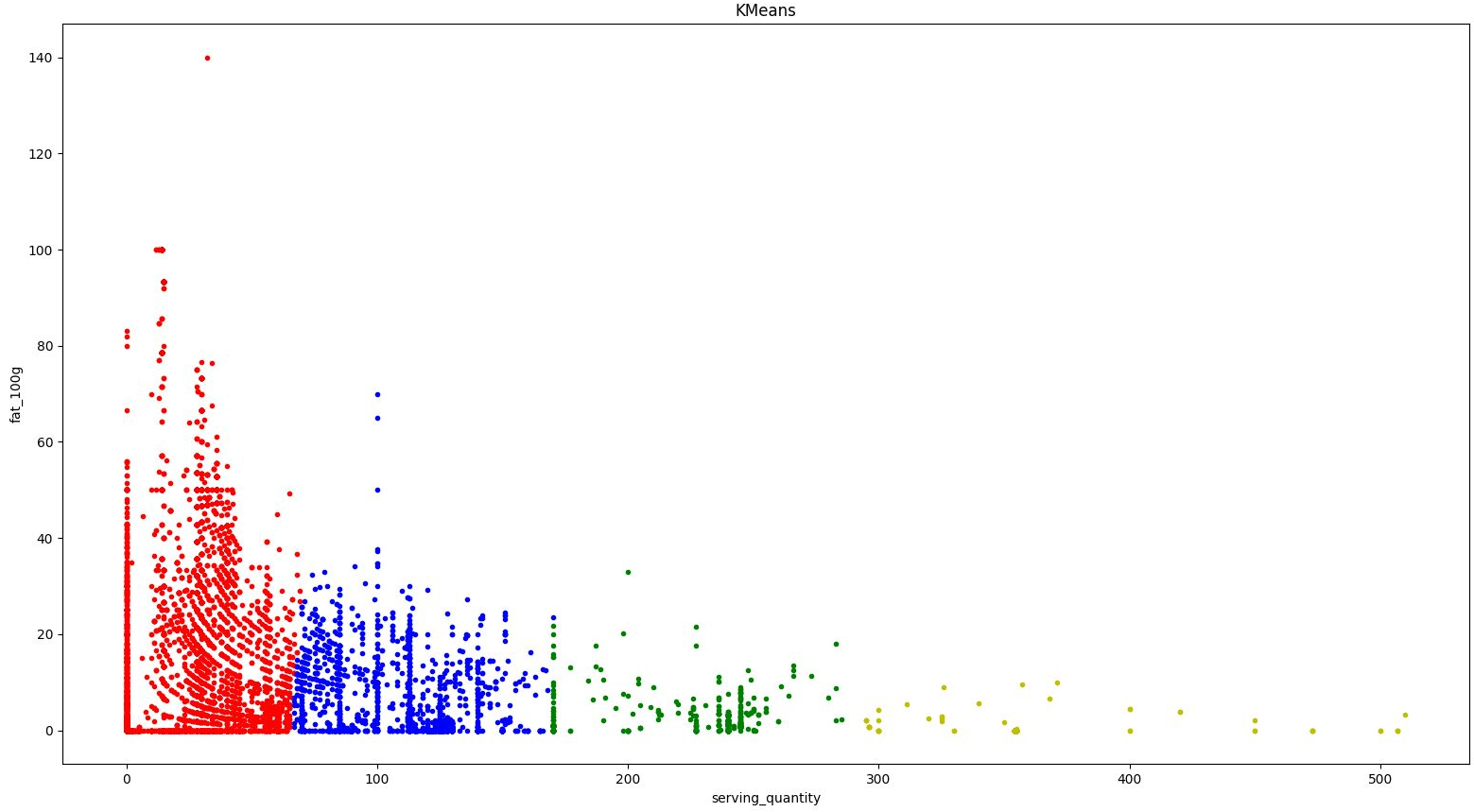
KMeans

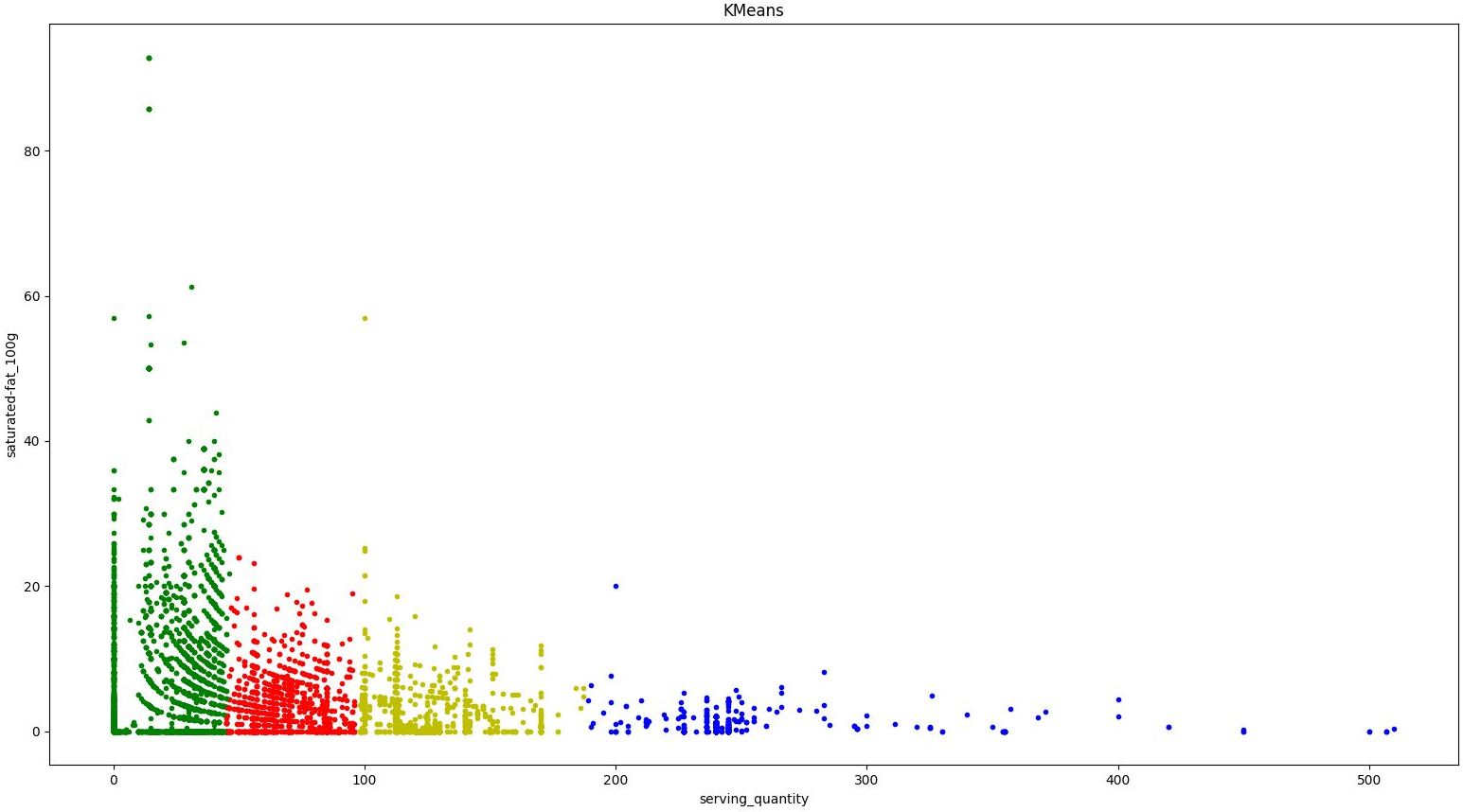

KMeans
carbohydrates_100g
proteins_100g
salt_100g
0
20
40
60
80
100
0
20
40
60
80
100
0
200
400
600
800

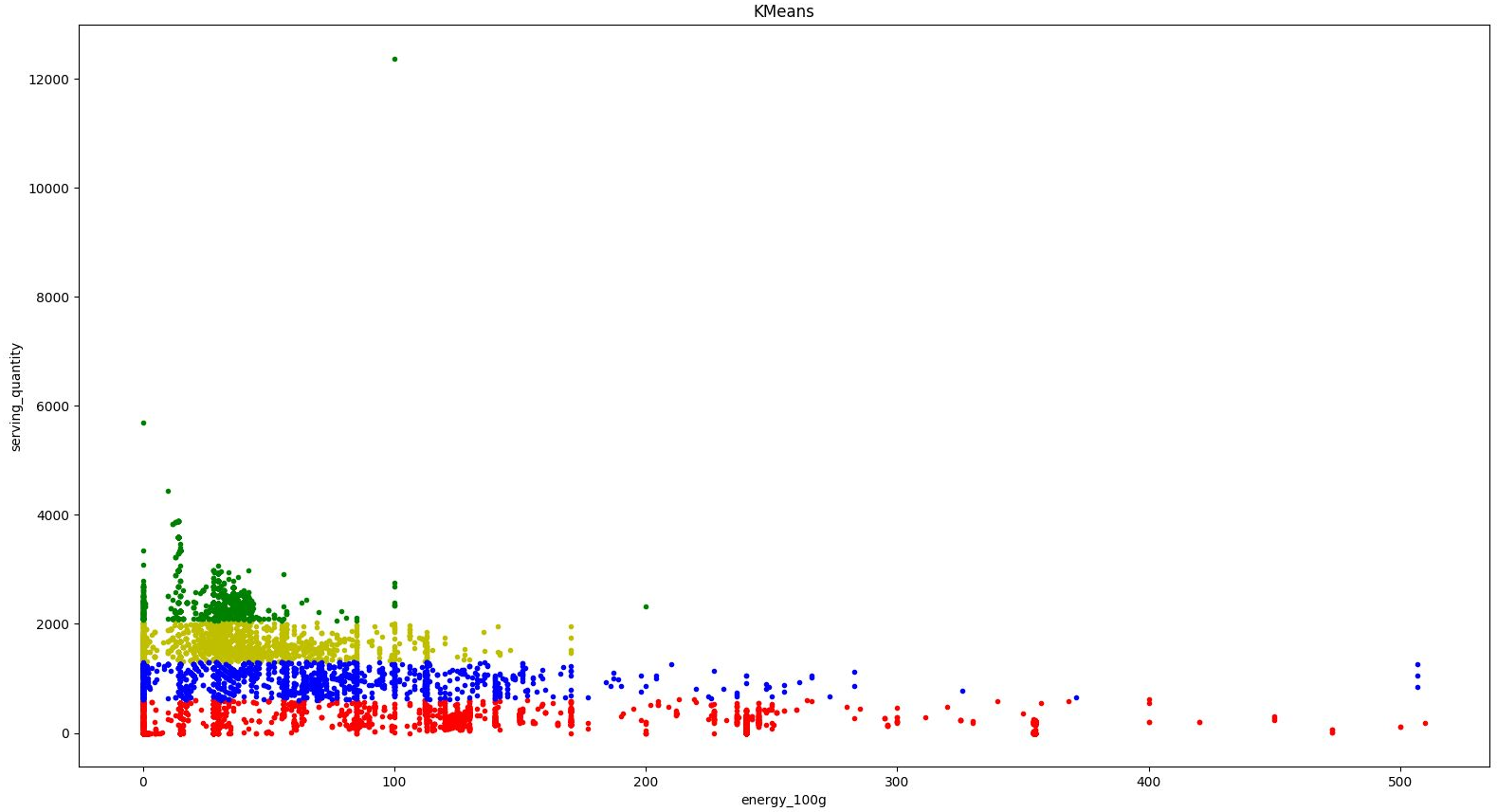
Conclusions et Améliorations possibles

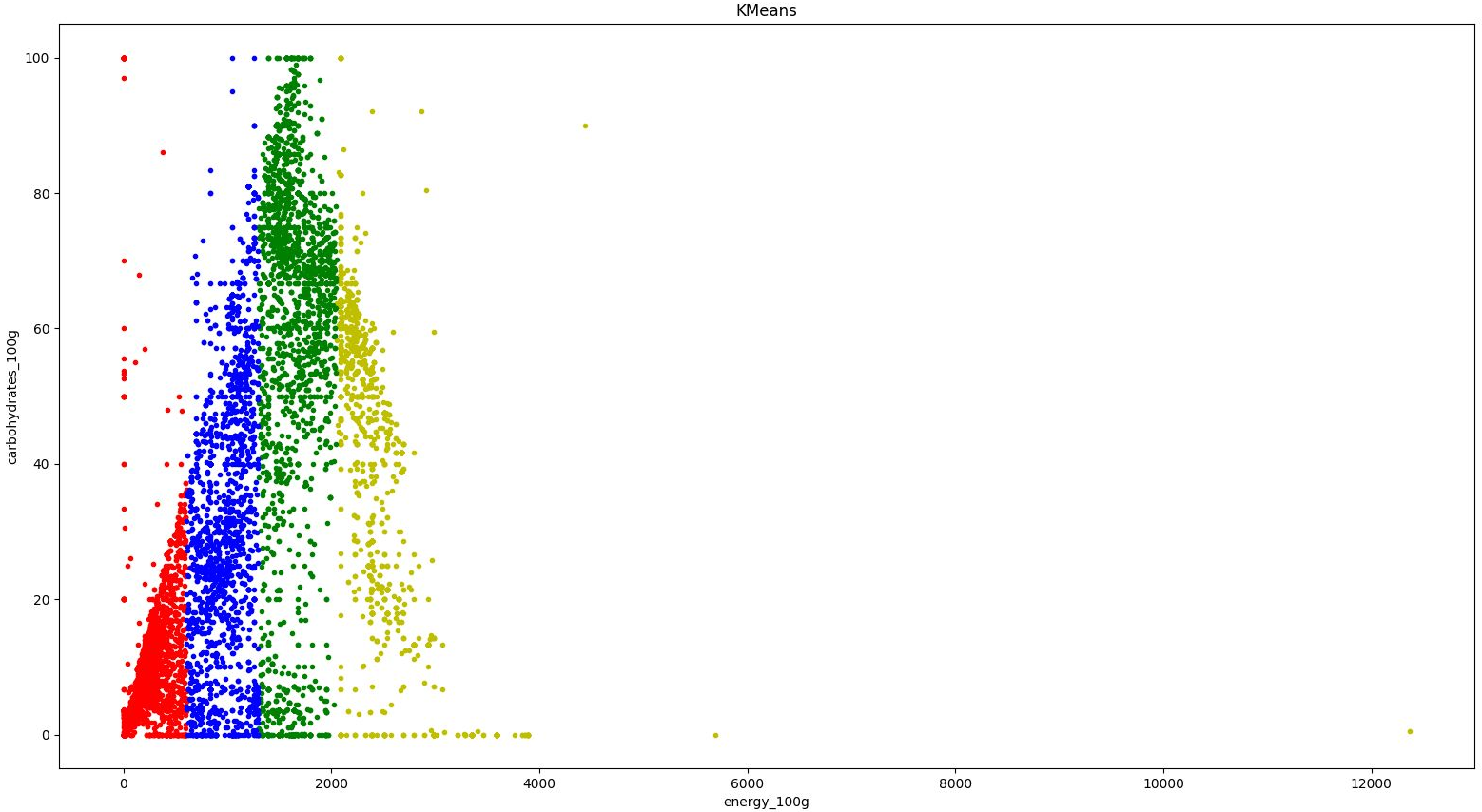
- Un ensemble de données très lacunaire, ce qui limite les analyses possibles
- On peut faire des classes d'aliments avec KMeans
- Les ACP sur les valeurs nutritionnelles nous montrent des données non fortement corrélées donc on a de la variété dans les aliments
- Vérification de la qualité du jeu de données
- Le pays de vente peut influer sur la composition des produits vendus
- Les analyses sont souvent longs sur une machine : utilisation d'un MapReduce pour paralléliser les opérations

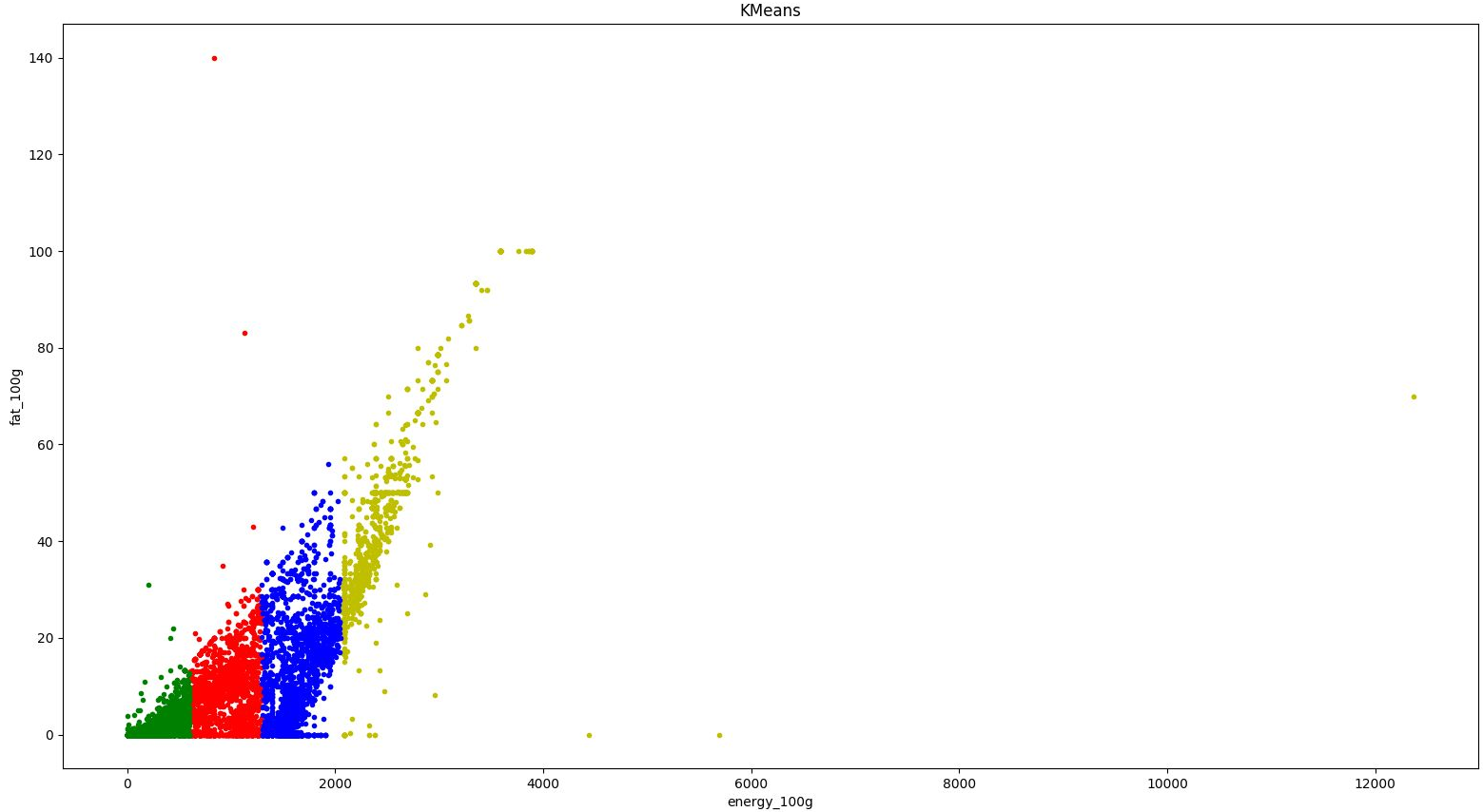
Annexes: Quelques graphes

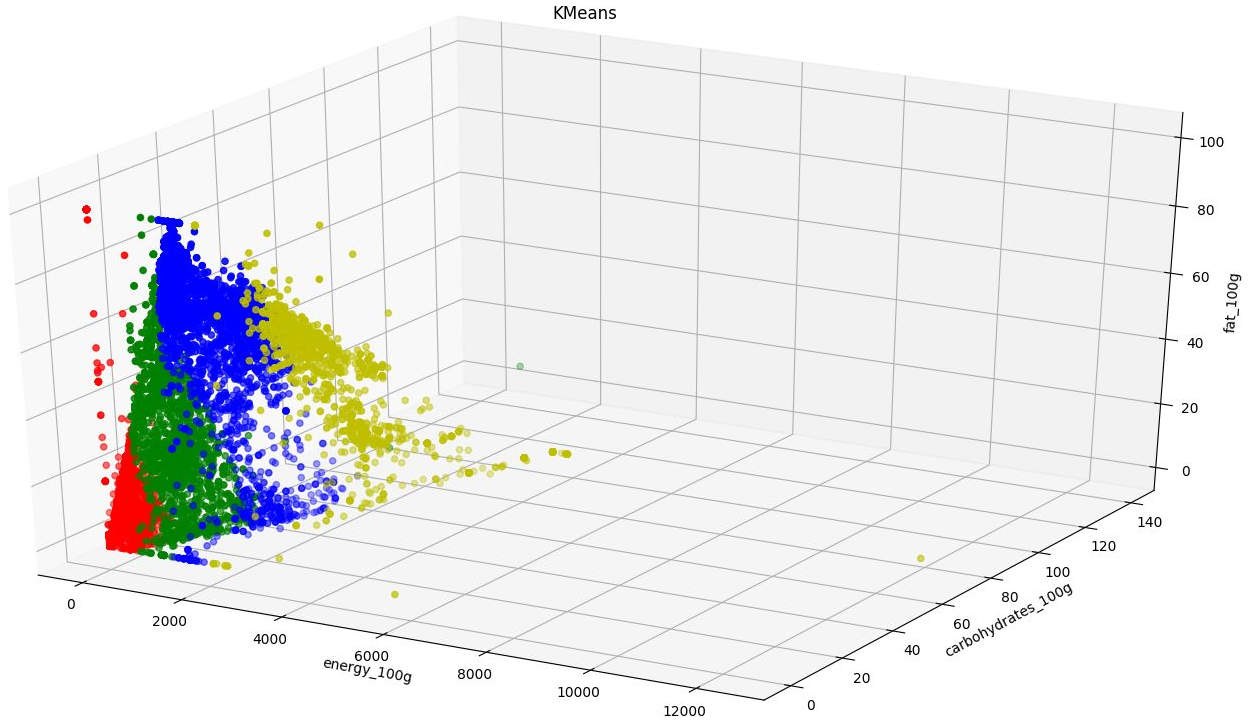

KMeans
fat_100g
serving_quantity
0
20
40
60
80
100
120
140
0
100
200
300
400
500

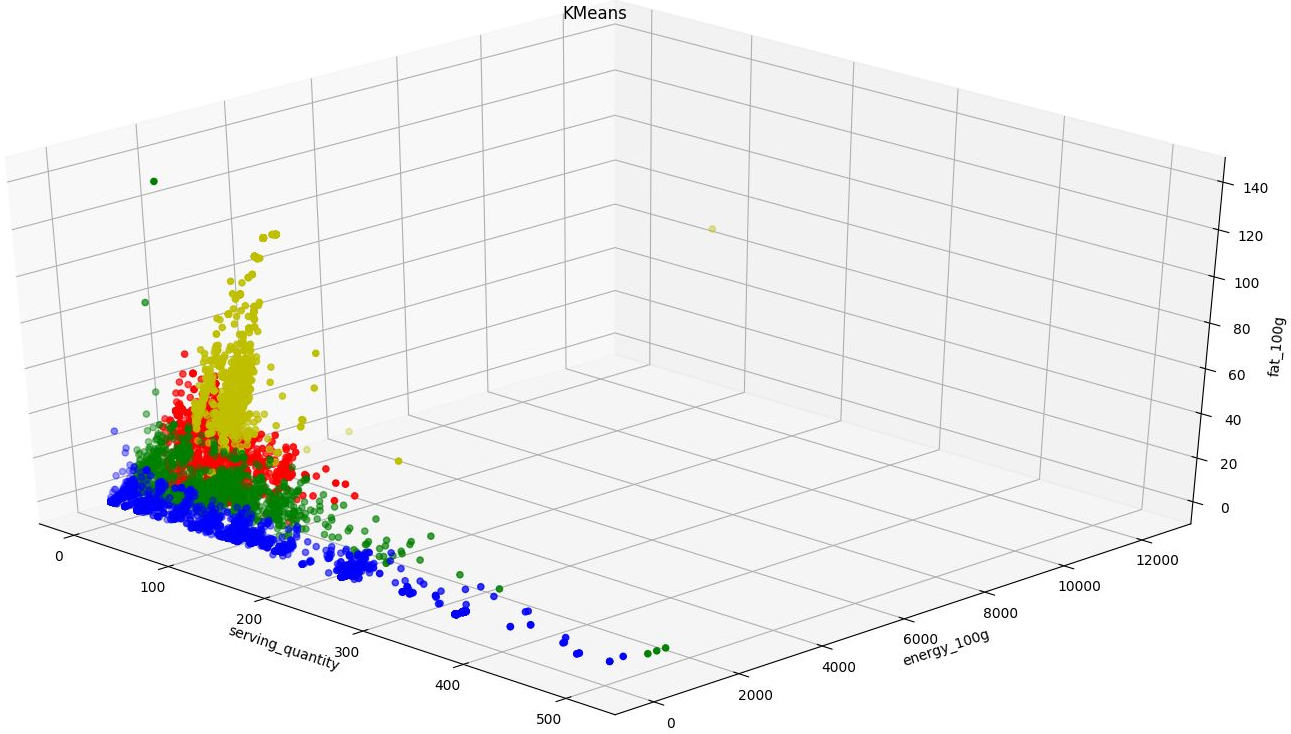

KMeans
saturated-fat_100g
serving_quantity
0
20
40
60
80
0
100
200
300
400
500

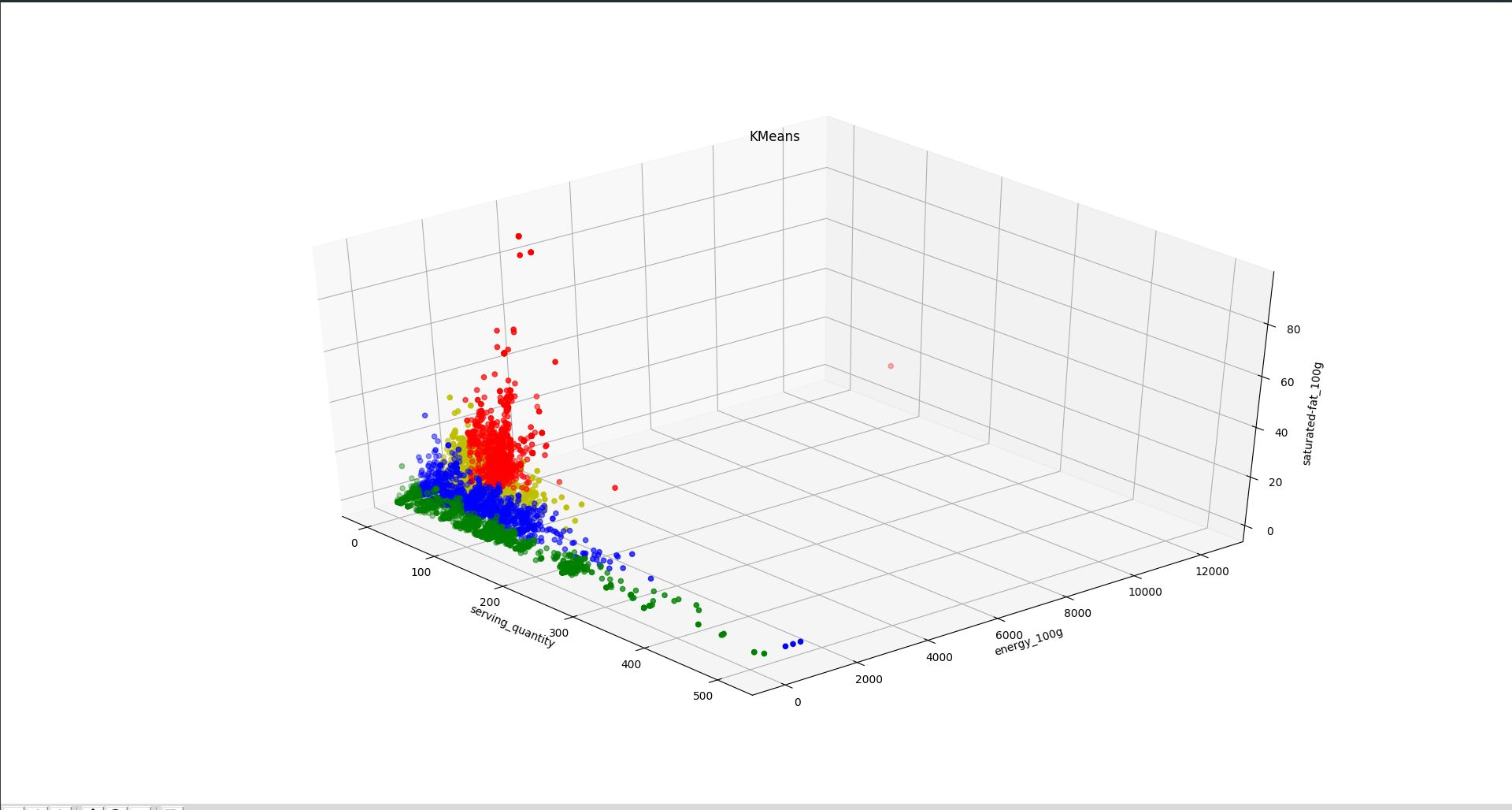

KMeans
serving_quantity
energy_100g
0
2000
4000
6000
8000
10000
12000
0
100
200
300
400
500


KMeans
carbohydrates_100g
energy_100g
0
20
40
60
80
100
0
2000
4000
6000
8000
10000
12000


KMeans
fat_100g
energy_100g
0
20
40
60
80
100
120
140
0
2000
4000
6000
8000
10000
12000


KMeans
energy_100g
carbohydrates_100g
fat_100g
0
2000
4000
6000
8000
10000
12000
0
20
40
60
80
100
120
140
0
20
40
60
80
100


KMeans
serving_quantity
0
100
200
300
400
500
energy_100g
0
2000
4000
6000
8000
10000
12000
fat_100g
0
20
40
60
80
100
120
140


KMeans
serving_quantity
0
100
200
300
400
500
energy_100g
0
2000
4000
6000
8000
10000
12000
saturated-fat_100g
0
20
40
60
80

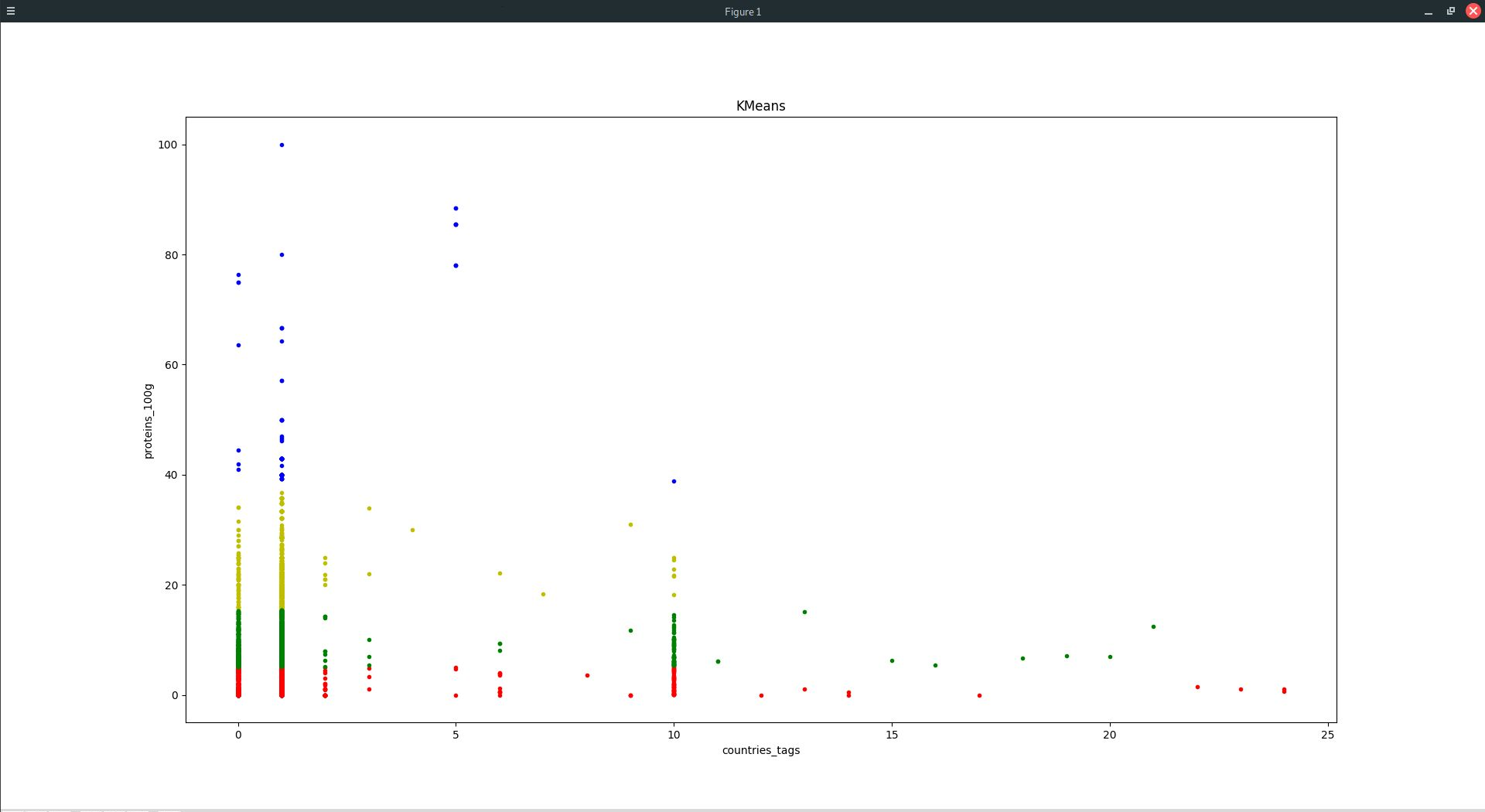
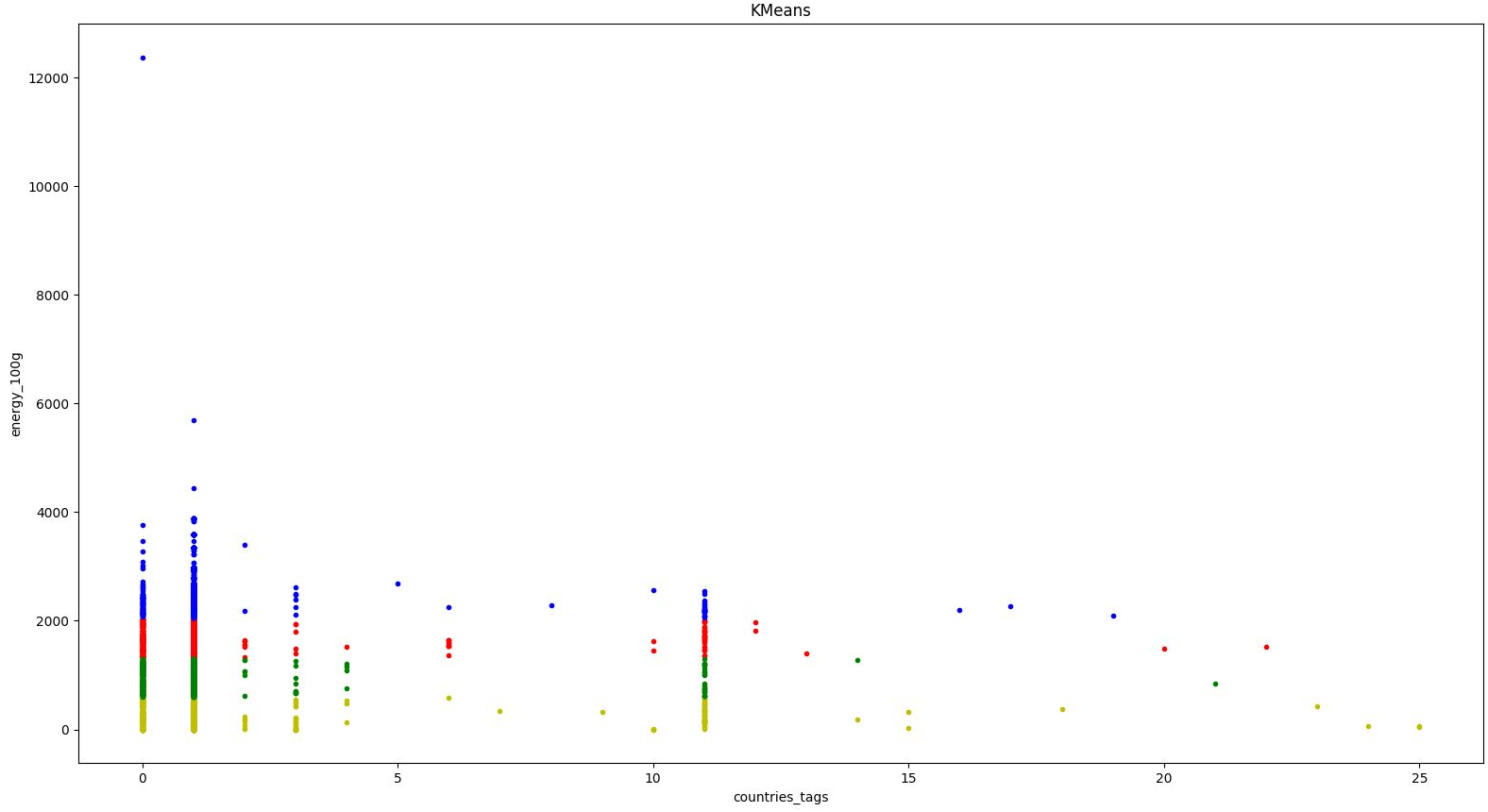


Figure 1
KMeans
proteins_100g
countries_tags
0
20
40
60
80
100
0
5
10
15
20
25


KMeans
energy_100g
countries_tags
0
2000
4000
6000
8000
10000
12000
0
5
10
15
20
25

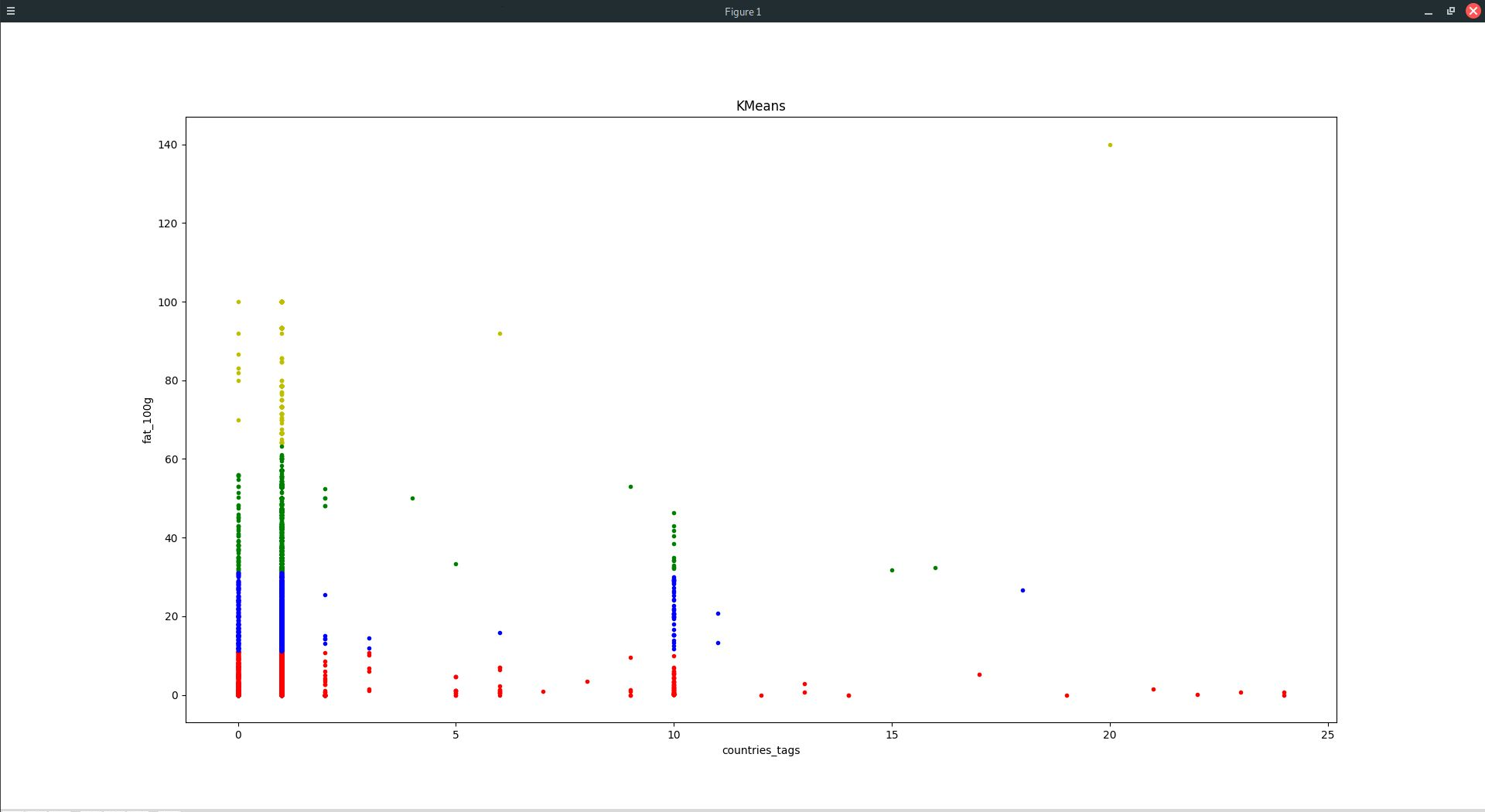
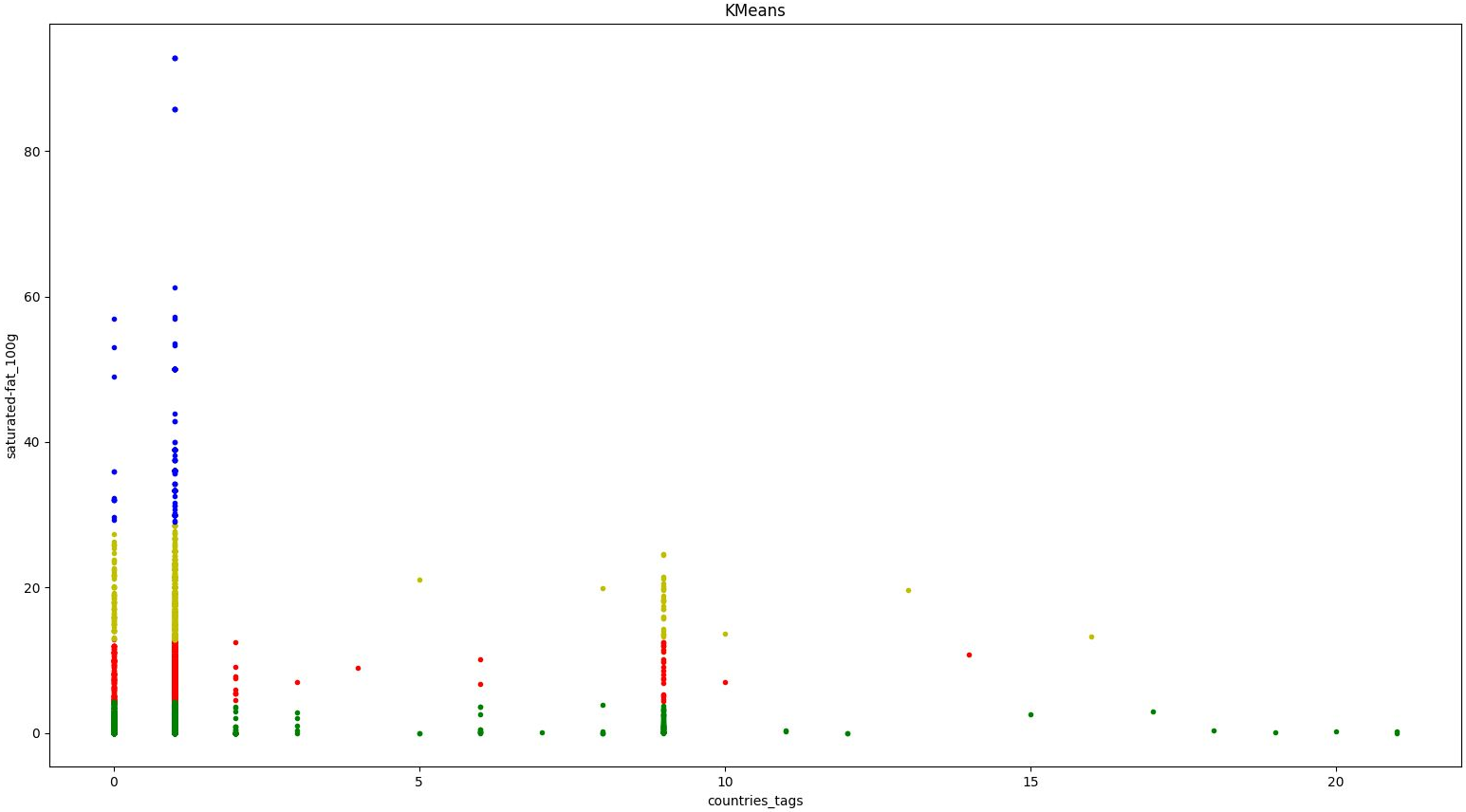


Figure 1
KMeans
fat_100g
countries_tags
0
20
40
60
80
100
120
140
0
5
10
15
20
25


KMeans
saturated-fat_100g
countries_tags
0
20
40
60
80
0
5
10
15
20