Agrupamento de Alimentos por Características Nutritivas

Semar Augusto da Cunha Mello Martins

Introdução:

Com o crescimento do mundo “fitness” as pessoas estão se tornando mais preocupadas com a alimentação. Seres humanos geralmente agrupam alimentos em 6 grupos principais, frutas, vegetais e legumes, grãos, proteínas, laticínios e comidas açucaradas. Será que esses grupos possuem macro-nutrientes e micro-nutrientes similares, sendo possível agrupá-los a partir somente dessas características ou será que os agrupamentos populares são totalmente arbitrários e não condizem com a qualidade da comida em si?

Na tentativa de responder essa pergunta, esse trabalho foi feito a partir da análise de um banco de dados com 292.415 entradas de alimentos.

2. Base de Dados:

A base de dados utilizada foi “Open Food Facts - Explore nutrition facts from foods around the world”. Contudo, foram usados somente as seguintes features da base de dados, uma vez que muitas features não são relevantes para a resposta da pergunta proposta.

Atributos:

- product_name - nome do produto
- energy_100g - quantidade de calorias em 100g
- fat_100g - quantidade de gordura em 100g
- saturated_fat_100g - quantidade de gordura saturada em 100g
- trans_fat_100g - quantidade de gordura trans em 100g
- carbohydrates_100g - quantidade de carboidratos em 100g
- fiber_100g - quantidade de fibras em 100g
- proteins_100g - quantidade de proteínas em 100g
- salt_100g - quantidade de sal em 100g
- nutrition-score-fr_100g - Avaliação feita pelo governo francês sobre a qualidade da comida. Será usado como critério de avaliação dos clusters
- pnns_groups_1 - A qual grupo alimentar o produto se encaixa
- pnns_groups_2 - A qual grupo alimentar o produto se encaixa

Dos quais product_name, nutrition-score-fr_100g e pnns_groups_1 e 2 não serão usados para fazer o agrupamento de fato.

O nutrition-score-fr_100g é um score tal que quanto menor o valor dele, melhor a qualidade da comida. A avaliação leva em consideração a quantidade de calorias, de gorduras, de gorduras saturadas, de carboidratos, açúcares simples, sódio e proteinas. O texto explicando como é avaliado está linkado nas referências.

Os campos pnns_groups_1 e pnns_groups_2 são campos preenchidos por usuários que correspondem a qual grupo o usuário pensa que o produto se encaixa. Os usuários colocaram por volta de 20 diferentes grupos nesses campos. Para a avaliação, foram agrupados alguns desses grupos em grupos maiores, por exemplo: “Biscuits and cakes” e “Chocolate products” foram ambos colocados em um grupo chamado “Sweets”.

A intenção era usar os micronutrientes (vitaminas e minerais) também. Contudo, a base de dados é extremamente mal preenchida e, ao se retirar os valores faltantes sobravam 0 produtos. Por esse motivo foram escolhidos somente os elementos principais de um alimento, ou seja, as calorias, os macronutrientes, a quantidade de gorduras maléficas ao nosso corpo, fibras e sódio (sal).

Apesar de só termos 12 features (de 162), ao retirarmos os valores faltantes o número de produtos caiu de 292.415 para 70.419 (24%)

product_name	pnns_groups_1	pnns_groups_2	energy_100g	fat_100g	saturated-fat_100g	trans-fat_100g	carbohydrates_100g	fiber_100g	proteins_100g	salt_100g	nutrition-score-fr_100g
Chaussons tress<U+00E9>s aux pommes	Sugary snacks	Biscuits and cakes	1090	10.700	2.000	0.667	38.70	2.000	3.330	0.647000	9
Pain Burger Artisan	unknown	unknown	1160	1.110	0.333	0.000	53.30	2.220	10.000	1.520000	1
Quiche Lorraine	Composite foods	Pizza pies and quiche	478	6.790	2.860	0.000	7.86	0.357	5.360	0.499000	2
Sea Salt Potato Chips	unknown	unknown	2243	32.140	3.570	0.000	57.14	3.600	7.140	0.952500	8
Clam Chowder A Condensed Soup	Composite foods	One-dish meals	276	1.640	0.000	0.000	9.02	0.800	4.100	1.249680	2

Amostra da Base de dados com os atributos selecionados:

A função cascadeKM, da biblioteca vegan, diz que a melhor divisão em grupamentos é usando 6 clusters. Isso induz que talvez a divisão tradicional talvez faça sentido.

3. Preparação dos Dados

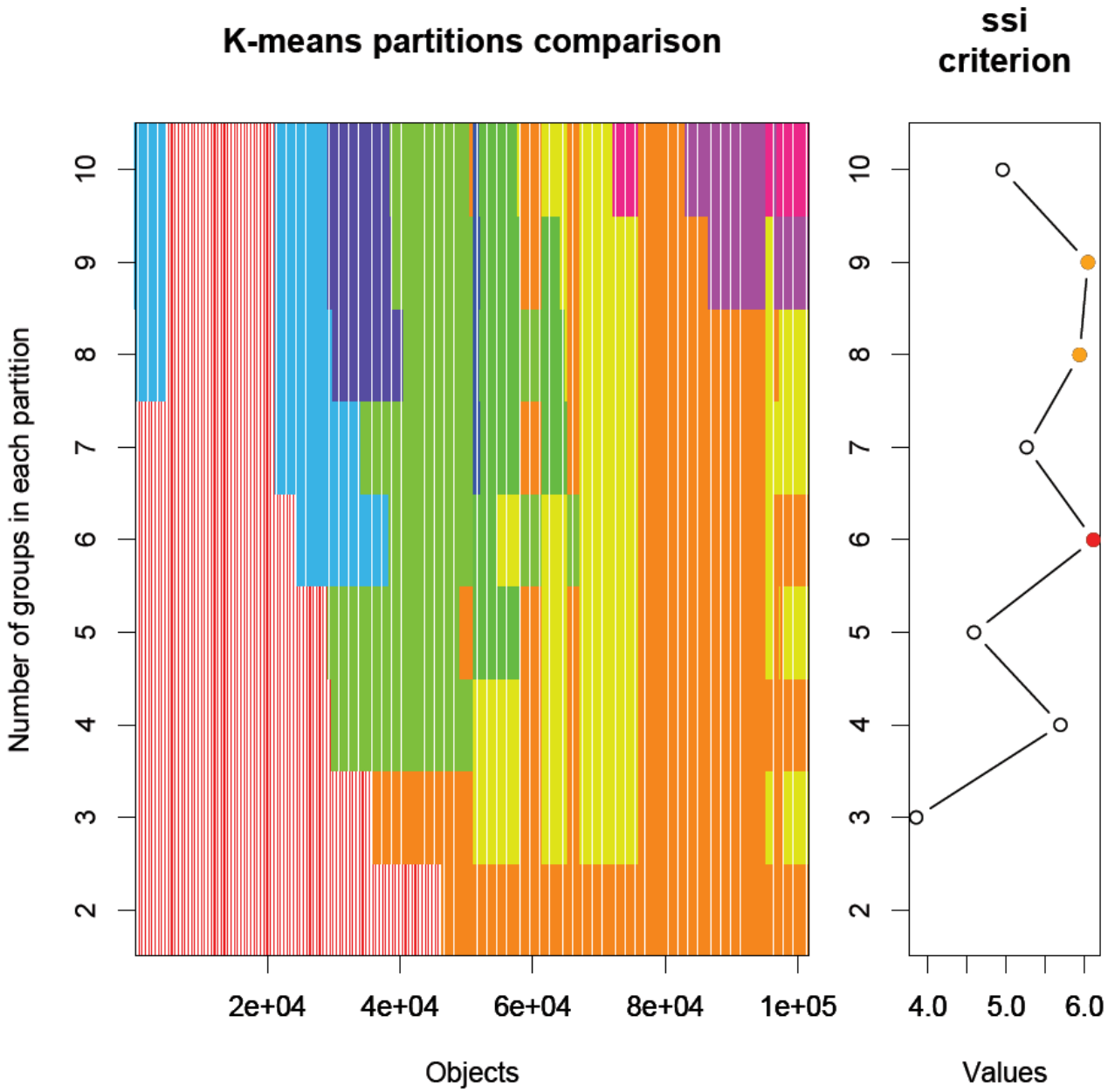
Durante a fase de preparação dos dados somente foi feita uma seleção dos produtos que não possuíam valores faltantes, pois os valores dos macronutrientes e micronutrientes usados não devem ser alterados para evitar perda de informação.

4. Modelagem

A modelagem do problema supôs que cada feature usada no cluster era um ponto em R^n e assumiu que kmeans conseguiria dividir nos clusteres esperados. Foram usadas as seguintes bibliotecas da linguagem R

- tidyverse - para manipulação dos dados
- vegan - para ver o número ideal de clusters
- cluster - para métodos de avaliação de qualidade dos clusters

5. Avaliação

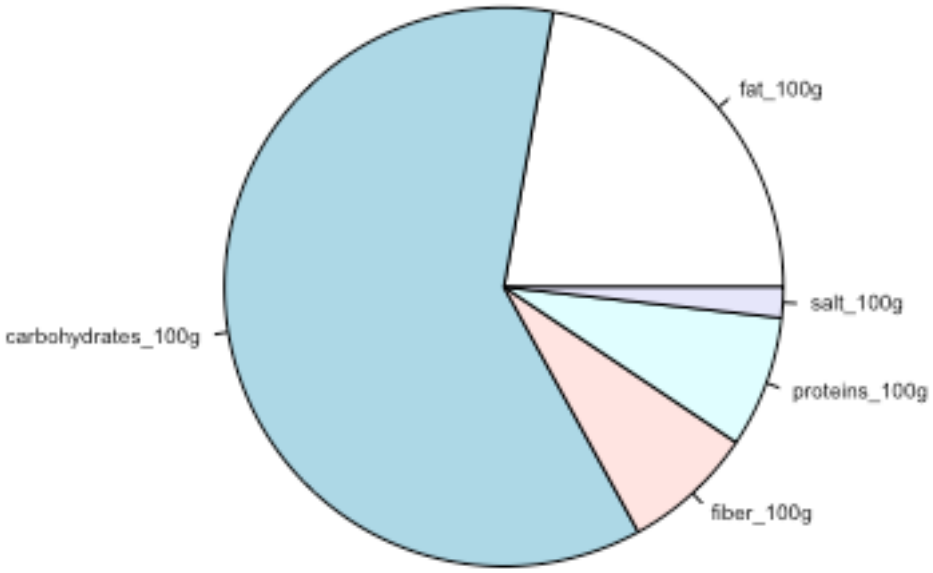


O algoritmo escolhido de agrupamento foi kmeans e a base de dados foi separada em 6 clusters em uma tentativa de comparar com os 6 grupos comumente usados por nós.

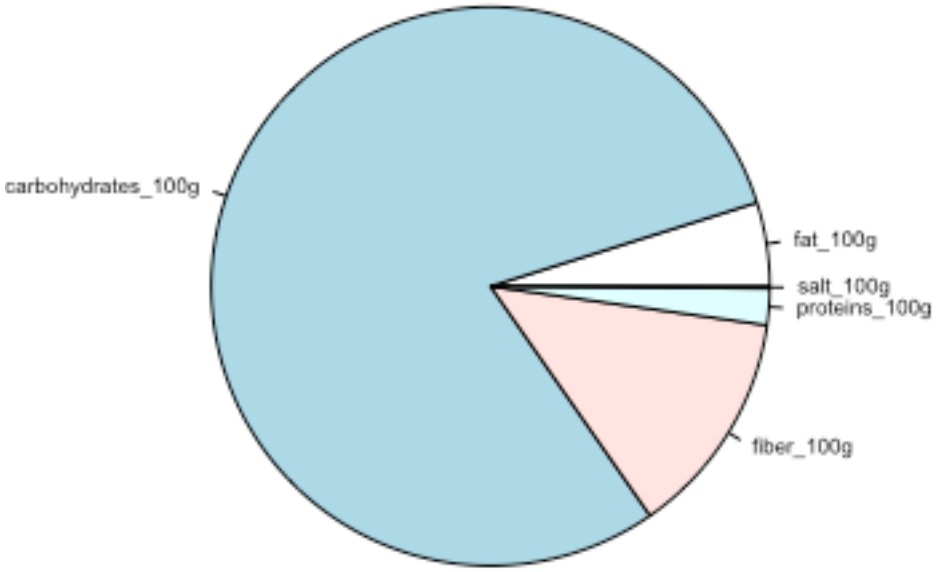
Todos os valores (exceto a quantidade de produtos em um cluster) correspondem à 100g da média de todos os produtos do grupo.

Começaremos por mostrar como é a composição dos grupos base

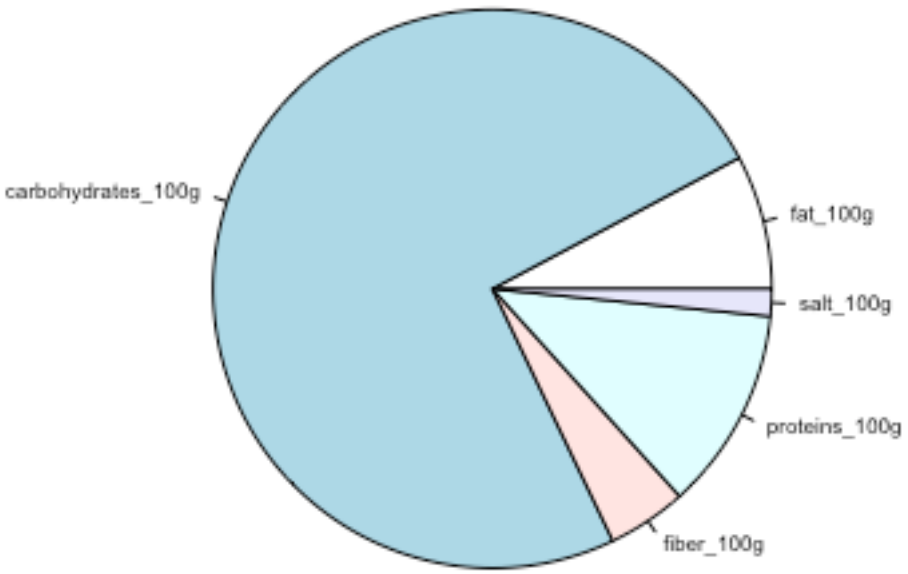
Vegetais:
- 240.81kcal
- 1.55g de gordura
- 8.51g de carboidratos
- 1.77g de proteínas
- 1.99g de fibras
- 0.48g de gordura saturada
- 0.00g de gordura trans
- 0.9g de sal
- Avaliação pelo governo Francês: -3.62



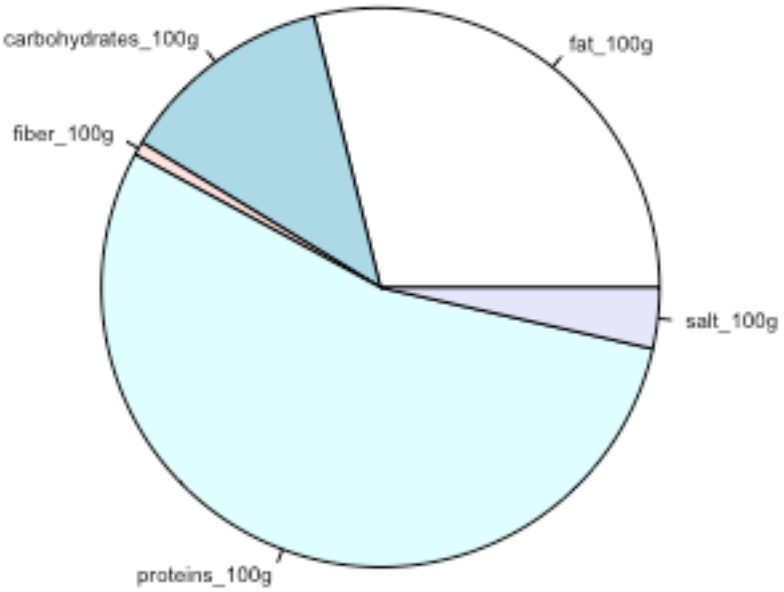
Frutas:
- 532kcal
- 1.6g de gordura
- 26.36g de carboidratos
- 0.69g de proteínas
- 4.32g de fibras
- 1.01g de gordura saturada
- 0.00g de gordura trans
- 0.03g de sal
- Avaliação do governo Francês: -1.46



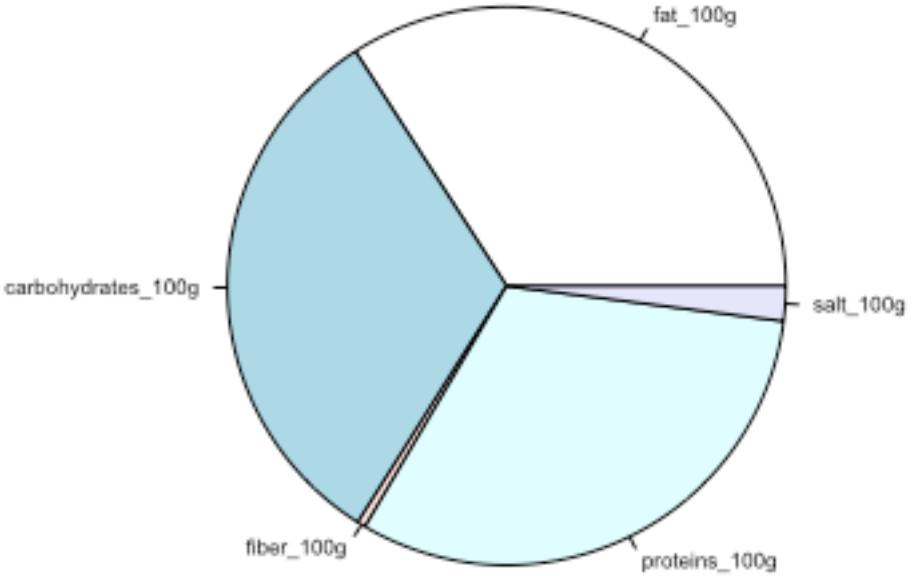
Grãos:
- 1408.85kcal
- 6.34g de gordura
- 60.58g de carboidratos
- 9.57g de proteína
- 3.74g de fibra
- 2.29g de gordura saturada
- 0.01g de gordura trans
- 1.28g de sal
- Avaliação do governo Francês: 3.65



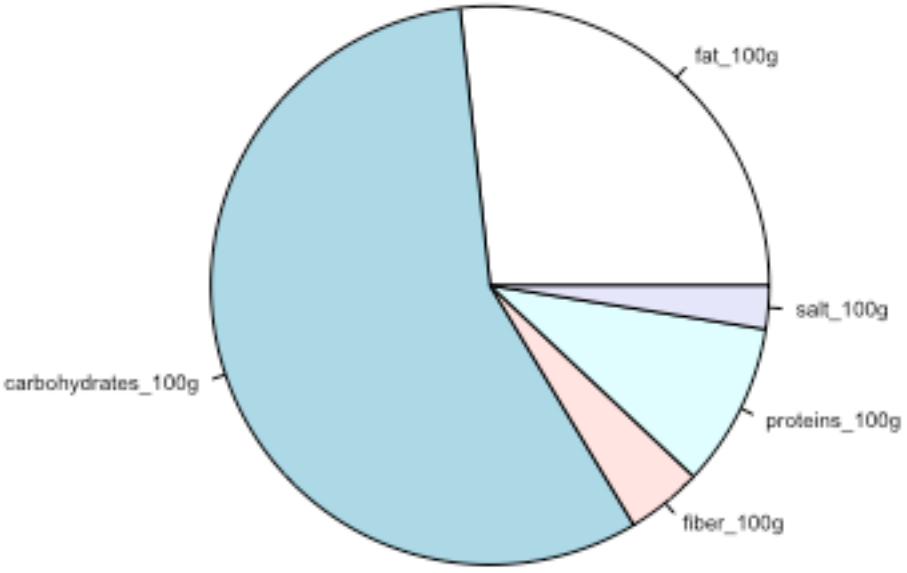
Proteínas:
- 749.51kcal
- 9.62g de gordura
- 4.15g de carboidratos
- 18.14g de proteina
- 0.26g de fibras
- 2.55g de gordura saturada
- 0.02g de gordura trans
- 1.17g de sal
- Avaliação do governo Francês: 3.93



Laticínios:
- 701.93kcal
- 10.46g de gordura
- 9.8g de carboidratos
- 9.64g de proteina
- 0.14g de fibra
- 6.3g de gordura saturada
- 0.04g de gordura trans
- 0.61g de sal
- Avaliação do governo Francês: 5.29

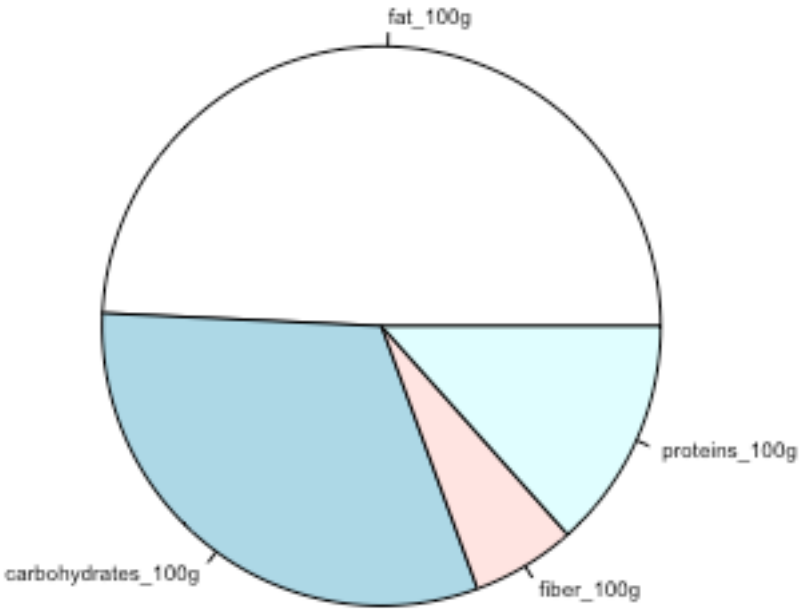


Comidas Açucaradas:
- 1515kcal
- 19.38g de gordura
- 41.25g de carboidratos
- 6.87g de proteína
- 3.18g de fibra
- 7.06g de gordura saturada
- 0.15g de gordura trans
- 1.81g de sal
- Avaliação pelo governo Francês: 12.64



Dados os grupos básicos, vamos mostrar os valores dos clusters

Cluster 1: 4561 produtos
- 2575,08kcal
- 49,46g de gorduras
- 31,45g de carboidratos
- 13,43g de proteínas
- 6,05g de fibras
- 13,62g de gorduras saturadas
- 0,10g de gorduras trans
- 1,33g de sal
- Avaliação pelo governo francês: 16,16



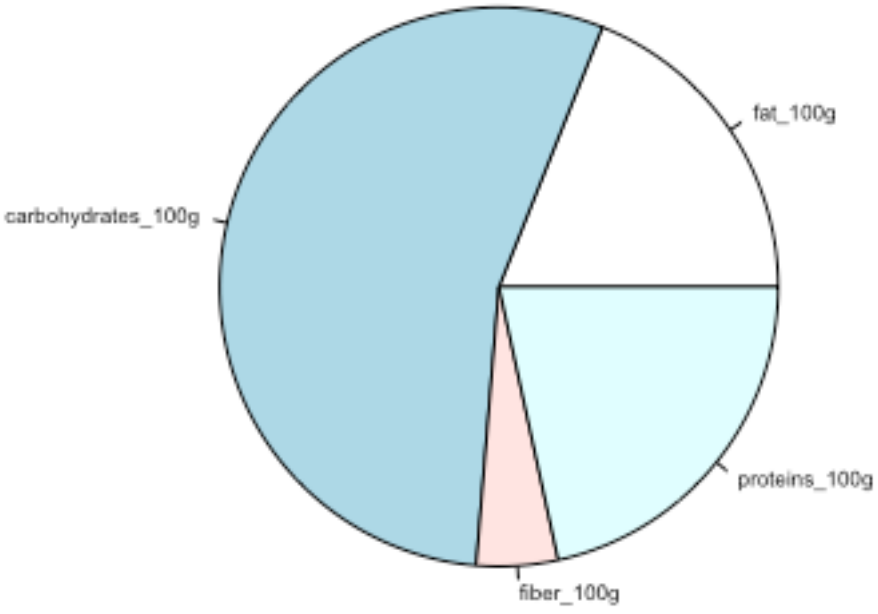
Amostra do cluster1:

product_name	energy_100g	fat_100g	saturated.fat_100g	trans.fat_100g	carbohydrates_100g	fiber_100g	proteins_100g	salt_100g	nutrition.score.fr_100g	grp.cluster
Peanuts, Mixed Nuts	2389	42.86	7.14	0.0	25.00	7.1	25.00	0.54356	14	1
Chocolate Scone	2912	39.29	23.21	0.0	80.36	3.6	5.36	0.63500	25	1
Original Buttery Spread	2690	71.43	10.71	0.0	0.00	0.0	0.00	1.81356	25	1
Buttery Spread, With Flaxseed Oil	2389	57.14	10.71	0.0	0.00	0.0	0.00	1.36144	22	1
Solid Milk Chocolate	2318	32.14	19.64	0.0	58.93	1.8	7.14	0.20320	24	1
Madelaine Chocolate Company, Chocolate	2301	32.50	20.00	0.0	57.50	2.5	7.50	0.19050	23	1
Milk Chocolate Rose	2389	33.33	21.43	0.0	57.14	0.0	9.52	0.18034	27	1
White Chocolate	2406	40.00	25.00	0.0	50.00	0.0	7.50	0.25400	28	1
The Madelaine Chocolate Company, All Natural Dark ...	2314	39.47	23.68	0.0	44.74	10.5	10.53	0.00000	17	1
The Madelaine Chocolate Company, Solid Milk Chocol...	2301	32.50	20.00	0.0	57.50	2.5	7.50	0.19050	23	1
The Madelaine Chocolate Comapny, Solid Milk Chocol...	2314	31.58	21.05	0.0	57.89	2.6	7.89	0.20066	23	1
The Madelaine Chocolate Company, Solid Milk Chocol...	2314	31.58	21.05	0.0	57.89	2.6	7.89	0.20066	23	1

Percebemos que o cluster 1 possui comidas extremamente calóricas e com alta taxa de gordura. Não se encaixa em nenhum dos grupos, mas o grupo com o qual o cluster 1 mais se assemelha é o das comidas açucaradas.

Ao se fazer uma análise manual do cluster 1, seria possível dizer que esse cluster se refere a produtos a base de chocolate e leite.

Cluster 2
- 9479 produtos
- 683,48kcal
- 6,60g de gorduras
- 19,05g de carboidratos
- 7,51g de proteínas
- 1,63g de fibras
- 0,06g de gorduras saturadas
- 0,13g de gorduras trans
- 1,66g de sal
- Avaliação pelo governo



francês: 5,59

Amostra do Cluster 2:

product_name	energy_100g	fat_100g	saturated.fat_100g	trans.fat_100g	carbohydrates_100g	fiber_100g	proteins_100g	salt_100g	nutrition.score.fr_100g	grp.cluster
Quiche Lorraine	478	6.79	2.86	0.00	7.86	0.357	5.36	0.49900	2	2
Ice Cream, Vanilla	883	11.27	7.04	0.00	22.54	0.000	2.82	0.14224	12	2
French Onion Dip	837	16.67	8.33	0.00	6.67	0.000	3.33	1.44018	17	2
Wild Alaskan Pink Salmon	598	7.94	1.59	0.00	0.00	0.000	19.05	0.92710	1	2
Mac 'n Cheese	761	7.58	4.04	0.00	21.20	0.505	8.08	0.75700	4	2
Barbecue Sauce	665	7.50	0.80	0.00	67.50	1.800	1.10	1.28000	11	2
Half & Half	556	10.00	6.67	0.00	6.67	0.000	3.33	0.12700	5	2
Red Curry Spice	540	0.00	0.00	0.00	14.29	14.300	14.29	9.32434	9	2
Cool Beans, Catalina Bean Hummus	895	21.43	3.57	0.00	10.71	3.600	3.57	1.13284	2	2
Cool Beans, Red Pepper Hummus	895	21.43	3.57	0.00	10.71	3.600	3.57	0.95250	2	2
Nuovo Pasta, Organic Ravioli, Butternut Squash	795	2.50	1.00	0.00	24.00	2.000	8.00	0.33020	-3	2

Percebe-se que o cluster 2 é um grupo de alimentos relativamente saudáveis, contudo possui uma taxa de gorduras trans anormal.

O cluster 2 não se assemelha aos laticínios pois possui uma proporção muito maior de carboidratos do que de gorduras e possui muito mais fibra do que o grupo dos laticínios.

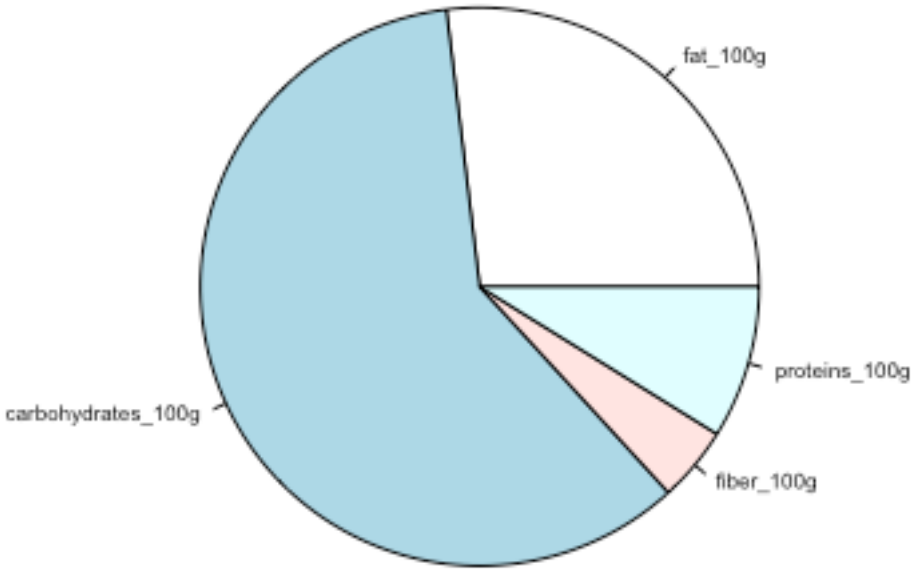
A quantidade de calorias e carboidratos do grupo faz com que ele se assemelhe mais com as frutas do que com os outros grupos, mas a taxa de gorduras saturadas, trans faz com que eles divirjam. a Avaliação francesa também mostra que eles não podem ser muito similares.

O conjunto dessas características mostra que o cluster 2 provavelmente é um grupo mais amplo do que qualquer dos 6 grupos base. Uma análise manual do cluster 2 nos dá a entender que o cluster 2 é um grupo extremamente genérico, nele há sorvetes, saladas, carnes, entre outros.

Cluster 3

- 13145 produtos
- 2019,18kcal
- 25,47g de gorduras
- 56,68g de carboidratos
- 8,37g de proteínas
- 4,11g de fibras
- 9,66g de gorduras saturadas
- 0,12g de gorduras trans
- 1,20g de sal

Avaliação média dada pelo governo francês: 17,45



Amostra do cluster 3

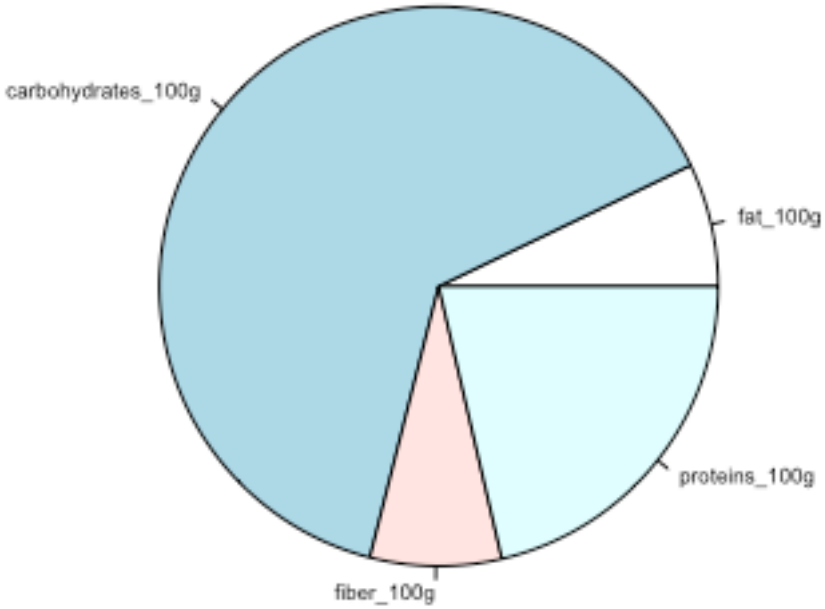
product_name	energy_100g	fat_100g	saturated.fat_100g	trans.fat_100g	carbohydrates_100g	fiber_100g	proteins_100g	salt_100g	nutrition.score.fr_100g
Banana Chips Sweetened (Whole)	2243	28.57	28.57	0.00	64.29	3.60	3.57	0.00000	14
Peanuts	1941	17.86	0.00	0.00	60.71	7.10	17.86	0.63500	0
Granola Cranberry And Acai	1824	10.91	0.91	0.00	74.55	5.50	9.09	0.25400	6
Banana Chips Sweetened	2243	28.57	28.57	0.00	64.29	3.60	3.57	0.00000	14
Milk Chocolate Pretzels	1883	22.50	12.50	0.00	70.00	2.50	5.00	1.01600	25
Scone	1920	25.88	15.29	0.00	52.94	2.40	3.53	0.41910	19
Belgian Vanilla Waffle	2067	22.35	9.41	0.00	67.06	2.40	8.24	1.70434	25
Seasonal Cookie Platter	1866	21.43	8.93	0.00	60.71	1.80	3.57	0.68072	20
Snickerdoodle Cookies	1866	21.43	8.93	0.00	60.71	1.80	3.57	0.99822	22
Freshly Baked Chocolat Croissant	1812	23.33	13.33	0.83	45.00	1.70	8.33	1.05918	19
Belgian Choc Chip Waffle	2117	22.35	9.41	0.00	68.24	2.40	8.24	1.70434	25
Piasten, Chocolate Assortment	2029	26.25	9.75	0.00	58.75	4.20	8.75	0.03048	11

O cluster 3 é outro grupo extremamente calórico, também se assemelha mais ao grupo das comidas açucaradas devido à taxa de gorduras e carboidratos. A avaliação dos franceses diz que o cluster 3 é ainda menos saudável do que o das comidas açucaradas, isso pode ter acontecido por causa de um número muito pequeno de produtos avaliados pelos usuários como “sweets”.

Uma análise manual no cluster 3 mostra que ele é um grupo bastante similar ao de “sweets”, ele abrange chocolates, biscoitos, pipocas doces, salgadinhos, entre outos.

Cluster 4

- 15100 produtos
- 237,60kcal
- 1,04g de gorduras
- 0,30g de gorduras saturadas
- 0,04g de gorduras trans
- 9,37g de carboidratos
- 1,11g de fibras
- 3,13g de proteínas
- 1,66g de sal
- Avaliação do governo francês: 1,17



Amostra do cluster 4:

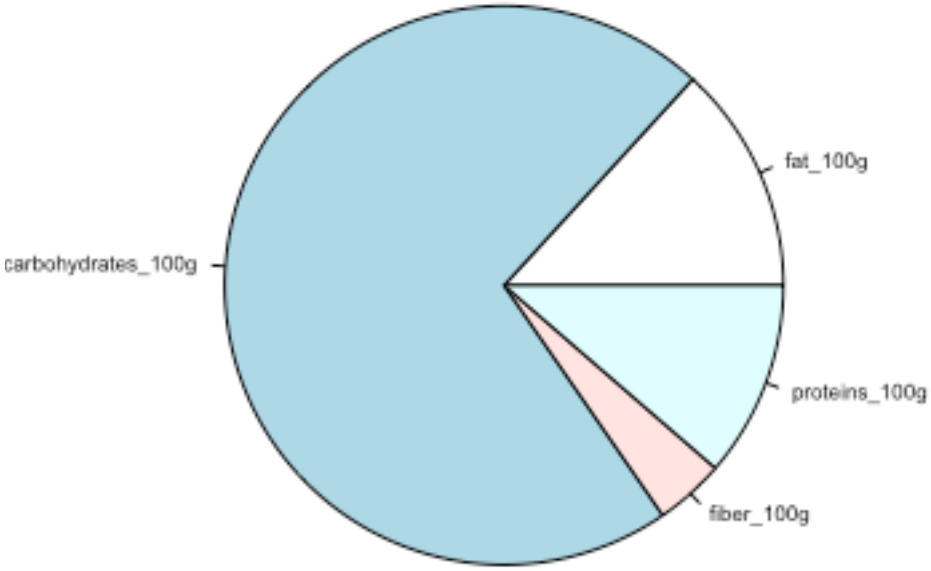
product_name	energy_100g	fat_100g	saturated.fat_100g	trans.fat_100g	carbohydrates_100g	fiber_100g	proteins_100g	salt_100g	nutrition.score.fr_100g
Fresh Organic Carrots	159	0.00	0.00	0.0	8.97	2.60	1.28	0.19558	-2
Romaine Hearts	75	0.00	0.00	0.0	3.53	2.40	1.18	0.01524	-3
Romaine	75	0.00	0.00	0.0	3.53	2.40	1.18	0.01524	-3
Green Leaf Lettuce	75	0.00	0.00	0.0	2.35	1.20	1.18	0.10414	-1
Cooking Spinach	100	0.00	0.00	0.0	3.53	2.40	2.35	0.19304	-4
Celery	59	0.00	0.00	0.0	3.64	1.80	0.00	0.26670	-1
Ryan's, Juice Melange, Fuji Pom Blend	192	0.00	0.00	0.0	11.67	0.00	0.00	0.03810	2
Ryan's, Lemonade	226	0.00	0.00	0.0	13.33	0.00	0.00	0.01016	2
Ryan's, Spiced Apple Cider	192	0.00	0.00	0.0	11.25	0.40	0.00	0.02540	2
Ryan's, Pink Lady, Apple Cider	192	0.00	0.00	0.0	11.67	0.00	0.42	0.01524	1

Percebemos que o cluster de número 4 é o cluster mais saudável dos criados. Ele se assemelha muito ao grupo dos vegetais. É possível afirmar que são grupos praticamente iguais dadas as caraterísticas dos alimentos.

Uma avaliação manual no grupo faz com que discordemos que o grupo é semelhante ao dos vegetais. De acordo com uma avaliação manual, o grupo na verdade corresponde a alimentos do mundo “fitness”. Consequentemente, boa parte do grupo é de vegetais, mas ele é mais abrangente que isso, possuindo clara de ovos, leite de amêndoa, sucos de frutas, cogumelos, peito de frango e etc.

Cluster 5

- 16783 produtos
- 1562,84kcal
- 11,04g de gorduras
- 4,65g de gorduras saturadas
- 0,07g de gorduras trans
- 59,74g de carboidratos
- 3,37g de fibras
- 9,49g de proteínas
- 1,43g de sal
- Avaliação do governo francês: 12,01



Amostra do cluster 5:

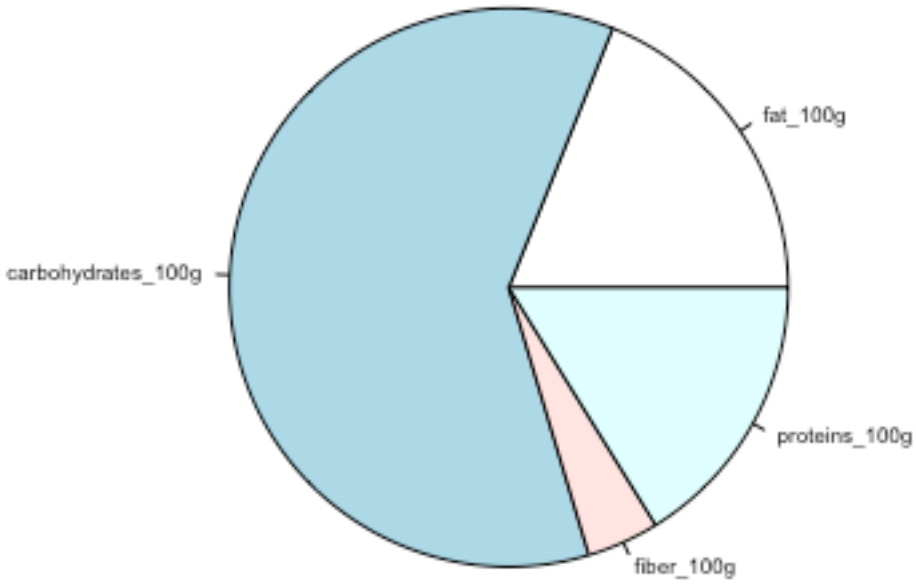
product_name	energy_100g	fat_100g	saturated.fat_100g	trans.fat_100g	carbohydrates_100g	fiber_100g	proteins_100g	salt_100g	nutrition.score.fr_100g
Granola Honey Almonds	1674	14.55	2.73	0.00	60.00	7.30	9.09	0.11430	0
Chili Mango	1569	2.50	0.00	0.00	87.50	2.50	2.50	1.96850	19
Butter Croissants	1523	16.88	10.39	0.00	44.16	1.30	7.79	1.08966	18
Wild Blueberry Muffins	1548	15.74	3.70	0.00	47.22	0.90	6.48	0.72898	14
Biscuit	1452	20.00	10.67	0.00	33.33	1.30	8.00	1.42240	19
Oatmeal Raisin Cookie	1778	17.70	8.85	0.00	61.95	1.80	3.54	0.56134	21
Sliced Plain Bagel	1527	1.18	0.00	0.00	71.76	2.40	12.94	1.19634	2
Muffin	1749	21.76	5.29	0.00	51.76	2.40	5.88	0.67310	16
Freshly Baked In Store Cherry Turnover	1674	22.35	11.76	0.00	45.88	1.20	4.71	0.77724	20
Apple Turnover	1623	22.35	11.76	0.00	42.35	2.40	4.71	0.80772	17

O Cluster 5, dadas as características médias dele, se assemelha ao grupo dos grãos, contudo, há uma taxa de gordura maior.
Em uma análise manual, percebe-se que o grupo é muito amplo e possui realmente muitos grãos, mas possui muitas sobremesas como cupcakes e cheesecakes. Acredito que a melhor caracterização do grupo seria chamá-lo de massas e sobremesas.

Cluster 6

- 11351 produtos
- 1133,89kcal
- 10,87g de gorduras
- 4,30g de gorduras saturadas
- 0,06g de gorduras trans
- 34,87g de carboidratos
- 2,40g de fibras
- 9,29g de proteínas
- 1,90g de sal

Avaliação do governo francês: 10,07



Amostra do cluster 6:

product_name	energy_100g	fat_100g	saturated.fat_100g	trans.fat_100g	carbohydrates_100g	fiber_100g	proteins_100g	salt_100g	nutrition.score.fr_100g
Cranberries	1255	0.00	0.000	0.000	83.33	10.000	0.00	0.00000	8
Turkish Apricots	1046	0.00	0.000	0.000	62.50	7.500	2.50	0.00000	8
Chaussons tress<U+00E9>s aux pommes	1090	10.70	2.000	0.667	38.70	2.000	3.33	0.64700	9
Pain Burger Artisan	1160	1.11	0.333	0.000	53.30	2.220	10.00	1.52000	1
Bolillos	1159	4.26	0.530	0.000	51.06	2.100	9.57	1.27000	1
Freshly Baked Italian Loaf	941	0.00	0.000	0.000	42.86	1.800	7.14	1.36144	1
Freshly Baked Apple Pie	1155	13.82	6.500	0.000	39.02	1.600	1.63	0.47498	14
La Brea Bakery Ciabatta Loaf Freshly Baked In Store	1176	3.51	0.000	0.000	52.63	1.800	8.77	1.78308	3
Lithuanian Rye Bread	954	2.63	0.000	0.000	47.37	1.800	5.26	1.42494	4
Organic Flourless Sprouted 7-Grain Bread	983	2.94	0.000	0.000	44.12	8.800	11.76	0.67310	-6

O cluster 6 possui um número de calorias alto (1133kcal/100g é um número alto), mas não tão alto quanto os clusters 1, 3 e 5. A quantidade de carboidratos indica uma semelhança ao grupo dos grãos, e isso se confirma com a presença de alguns tipos de massas, pizzas e pães.

Com base nisso e em uma análise manual, esse grupo poderia ser chamado de “derivados do trigo”. A maioria dos alimentos presentes no grupo são massas, pizzas e pães.

Conclusão:

Não é possível agrupar os alimentos nos 6 grupos tradicionais considerando somente as quantidades de proteínas, gorduras, carboidratos, fibras e sal. Ao considerarmos somente essas caraterísticas, conseguimos grupos mais similares a:

- “Alimentos à base de chocolates e leite”
- “Diversos”
- “Doces"
- “Alimentos do mundo fitness”
- “massas e sobremesas”
- “derivados do trigo”

Contudo, há uma sobreposição desses grupos entre si mesmos, não foi possível descobrir os o que causa que um produto esteja no cluster 6 e não no cluster 5 ou vice-versa.

Independente disso, foi possível perceber que os alimentos vendidos hoje em dia são extremamente diversos e não podem ser colocados em grupos tão claros como os tradicionais. Será que uma pizza seria colocada como um grão, já que é feita de trigo ou como um laticínio, já que uma quantidade significativa de suas calorias vem do queijo que está nela? Um cogumelo seria um vegetal? E uma barrinha de proteína, será que ela pertence ao grupo das “proteínas” ou dos “doces”?

Referências:

https://www.kaggle.com/openfoodfacts/world-food-facts

https://fr.openfoodfacts.org/files/hcspa20150625_infoqualnutprodalim.pdf

https://rstudio-pubs-static.s3.amazonaws.com/33876_1d7794d9a86647ca90c4f182df93f0e8.html

https://www.statmethods.net/advstats/cluster.html