Análise de Componentes Principais (PCA)

Gustavo Pinho - Matheus Ileck

June 2023

1 Introdução a Análise de Componentes Principais (PCA)

.

A Análise de Componentes Principais (PCA) é uma técnica estatística amplamente utilizada para redução de dimensionalidade e exploração de dados multivariados. Ela busca capturar a maior parte da variabilidade presente nos dados através de uma transformação linear dos atributos originais em um novo conjunto de componentes principais. Esses componentes principais são combinações lineares dos atributos originais, ordenados de forma que o primeiro componente principal explique a maior variação nos dados, o segundo componente principal explique a segunda maior variação, e assim por diante.

O PCA tem várias aplicações, como simplificação e visualização de dados complexos, detecção de padrões, redução de ruído, normalização de dados e compressão de informações. Ao reduzir a dimensionalidade dos dados, o PCA permite que sejam analisados e interpretados mais facilmente, além de ajudar na eliminação de redundâncias e na identificação das principais características que contribuem para a variabilidade dos dados.

Em resumo, o PCA é uma técnica estatística poderosa que permite extrair informações valiosas a partir de dados multivariados, reduzindo sua dimensionalidade e capturando as principais características. É uma ferramenta fundamental em áreas como ciência de dados, aprendizado de máquina, análise de dados e pesquisa, fornecendo insights e facilitando a compreensão dos dados de forma mais concisa e informativa.

2 PCA do cardápio do McDonalds

.

Utilizamos a técnica do PCA para fazer uma análise em cima do cardápio de lanches e bebidas oferecidas pela rede de Fast Food McDonald's ao redor do mundo. Dividimos o cardápio em 3 grupos: lanches salgados, lanches açucarados/sobremesas e, por último, as bebidas. Dentro desses grupos, iremos relacionar a correlação e, se possível, mostrar um padrão entre os valores de macro e micronutrientes de cada grupo. Os macro e micronutrientes que serão as colunas da nossa tabela são, respectivamente, o tamanho da porção, calorias, proteína, gordura total, gordura saturada, gordura trans, colesterol, carboidrato, açúcar, açúcar adicionado e sódio.

2.1 Introdução

.

Primeiro utilizamos o Kaggle para baixar um dataset comas diferentes comidas oferecidas pelo McDonalds ao redor do mundo. Logo após isso normalizamos e manipulamos os valores das tabelas já que em alguns campos havia valores muito discrepantes, no caso a coluna do sódio tinha valores muito mais altos

que os demais, e isso iria fazer com que o PCA tendesse mais para esse campo em específico, então dividimos os valores dessa coluna por mil.

sodium	sodium
1,075	1.075
1087.46	1087.46
1051.24	1051.24
1529.22	1529.22
579.6	579.6

Após isso jogamos nossa tabela no software Past para poder que ele faça o PCA e possibilite que possamos plotar os grafos de nossas tabelas.

	servesize	calories	protien	totalfat	satfat	transfat	cholestrol	carbs	sugar	addedsugar	sodium
McSpicy™ Paneer Burger	190	652	20.29	39.45	17.12	0.18	21.85	52.33	8.35	5.27	1.07458
Spicy Paneer Wrap	250	674	20.96	39.1	19.73	0.26	40.93	59.27	3.5	1.08	1.08746
American Veg Burger	177	512	15.3	23.45	10.51	0.17	25.24	56.96	7.85	4.76	1.05124
Veg Maharaja Macã	306	832	24.17	37.94	16.83	0.28	36.19	93.84	11.52	6.92	1.52922
Green Chilli Aloo Naan p	132	356	7.91	15.08	6.11	0.24	9.45	46.36	4.53	1.15	0.5796

A divisão dos produtos se deu da Seguinte forma, lanches salgados são os de cor preta, lanches açucarados/sobremesas são os de cor rosa, e por último, as bebidas que são as de cor azul clara. As cores irão ajudar a fazer a diferenciação dos três grupos alimentares no momento em que iremos plotar nossos dados.

American Triple Cheese Chicken	Double Chocochips Muffin
American Triple Cheese Veg	Vanilla Chocochips Muffin
Cheese Lava Burger	Veg McMuffin
Chicken Cheese Lava Burger	Double Cheese McMuffin
Chunky Chipotle American Burger Chick	Spicy Egg McMuffin

Strawberry Ice Tea
Green Apple Ice Tea
Iced Coffee
Cold Coffee Frappe
Small Coca-Cola

2.2 Breve descrição da plataforma Past

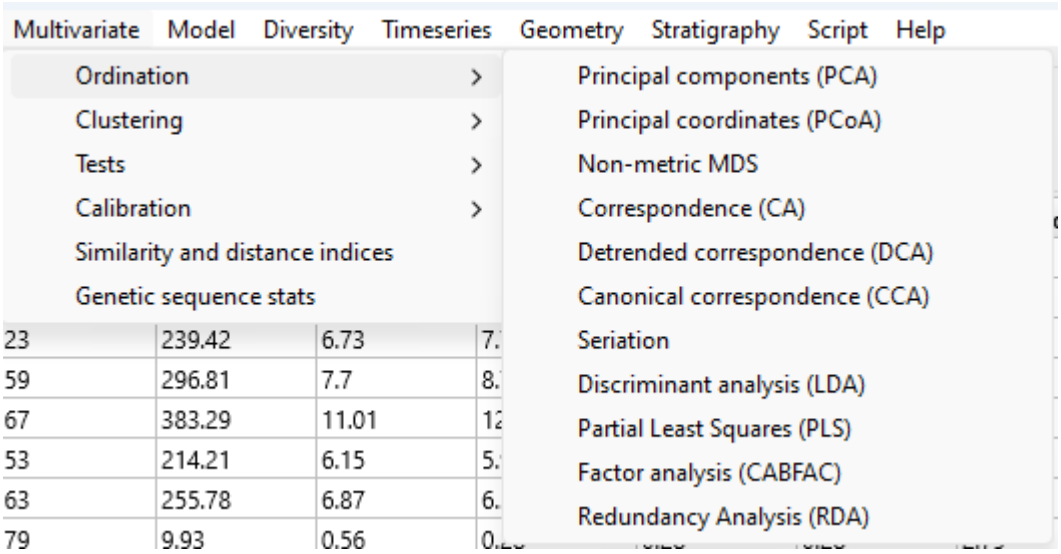
. Para plotar nossos gráficos decidimos usar o software Past, ele é um software estatístico gratuito de código aberto usado para análise de dados multivariados. Ele fornece uma ampla gama de recursos estatísticos e gráficos para explorar e visualizar dados complexos. O software é especialmente útil para análises

exploratórias de dados e análises de componentes principais (PCA), análises de correspondência, análises de regressão e muito mais.

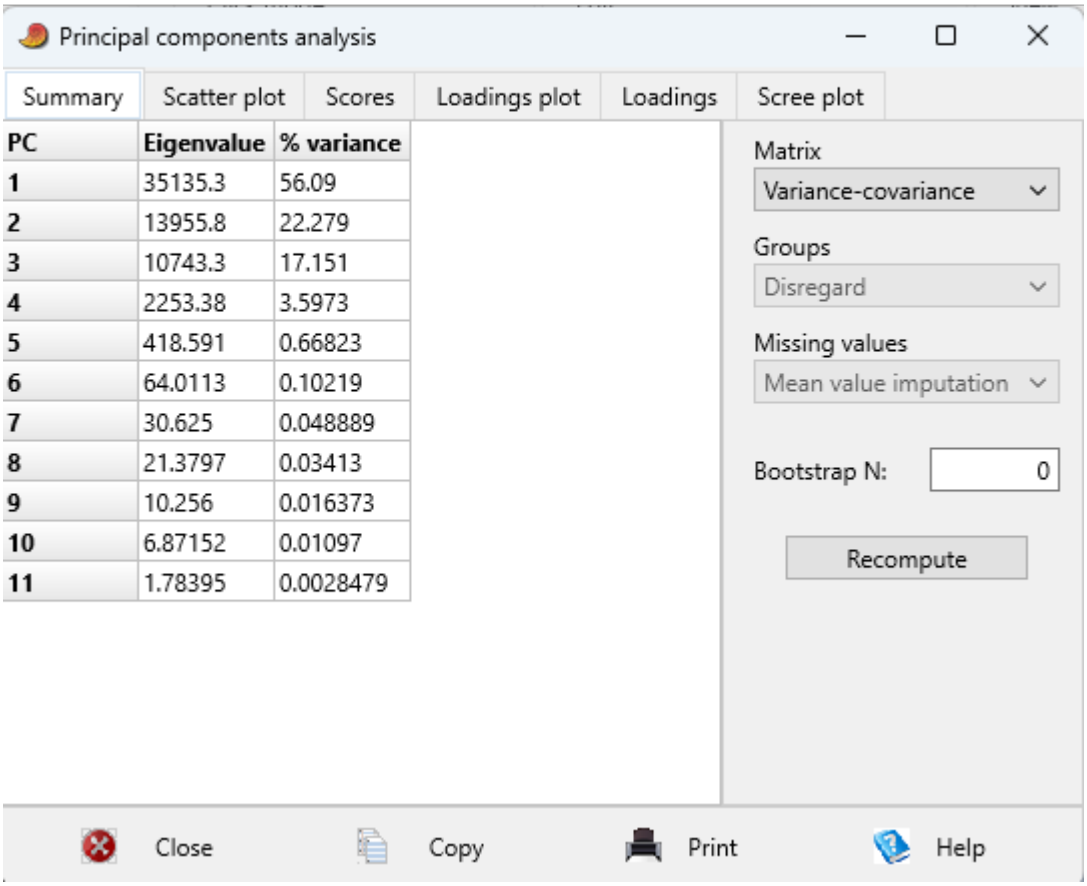
O Past oferece uma interface intuitiva e amigável, permitindo que os usuários importem dados de várias fontes, realizem análises estatísticas e produzam visualizações de alta qualidade. Ele suporta uma ampla variedade de gráficos, incluindo gráficos de dispersão, gráficos de barras, gráficos de linha, gráficos de superfície e mapas de calor, que ajudam a representar visualmente as relações entre variáveis.

Além disso, o Past também possui recursos avançados, como análise de agrupamento, análise discriminante, análise de sobrevivência, análise de redes, entre outros. Ele permite que os usuários personalizem suas análises e gráficos de acordo com suas necessidades, fornecendo opções de configuração e ajuste dos parâmetros.

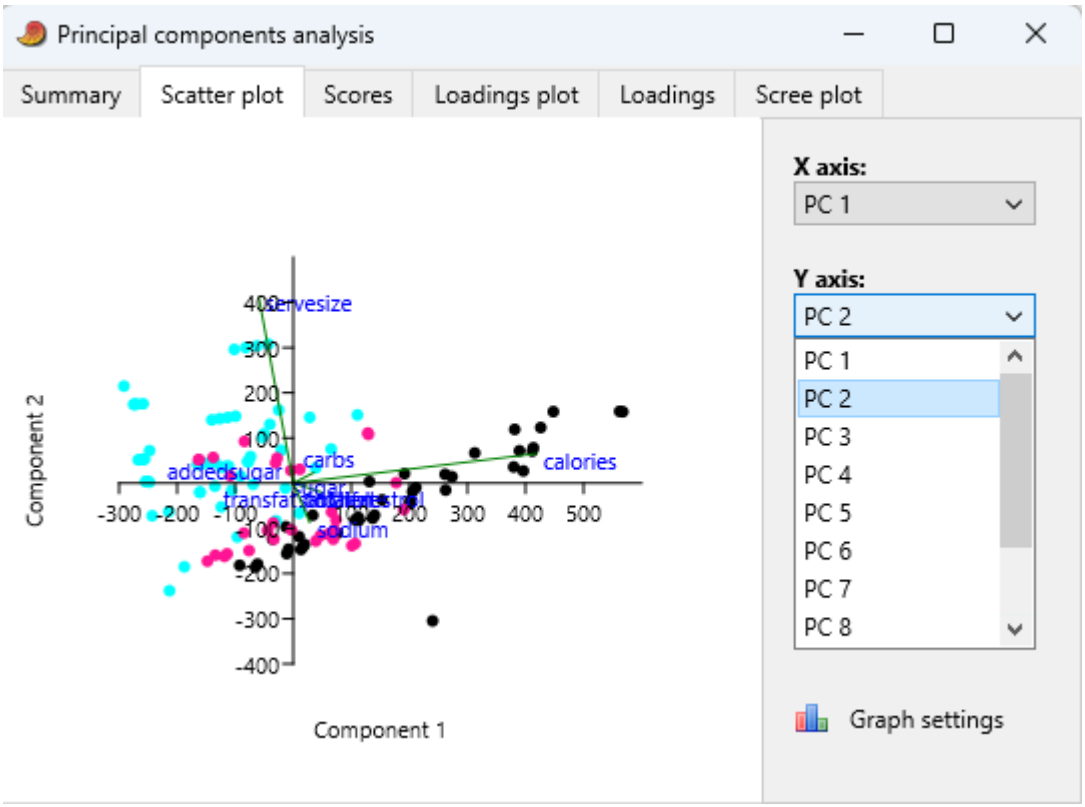
Como um software de código aberto, o Past é amplamente utilizado em pesquisas científicas, análises de dados biológicos, análises de ecologia, estudos de população, análises sociais e outras áreas que exigem análise e visualização de dados multivariados. Para conseguir plotar os resultados da tabela você deverá deixar o mouse por cima de Multivariate na parte superior, logo após isso você deverá selecionar Ordination e depois clicar sobre Principal Components (PCA).



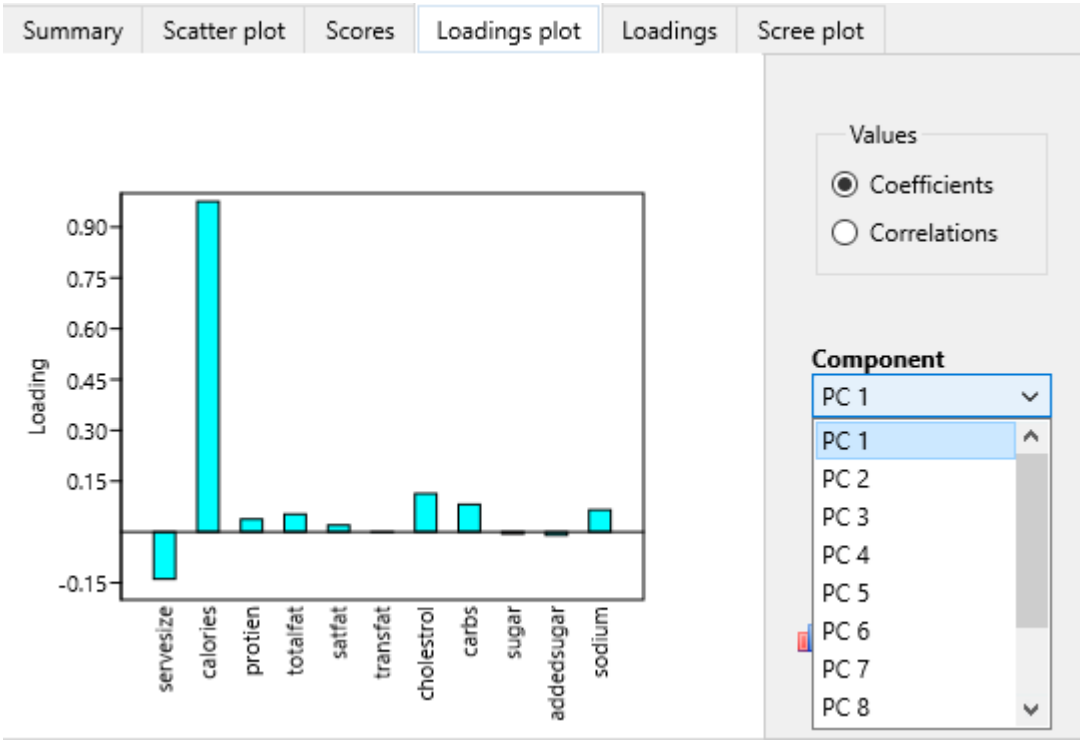
Logo após isso, irá abrir uma janela com uma tabela que terá 3 colunas, a primeira com os PCAs, a segunda com os autovalores, e a terceira com a porcentagem de variância.



Clicando em Scatter plot, temos os nossos dados já plotados com os eixos X e Y na lateral com os componentes principais que eles estão usando como base. Com isso poderemos ter um representação visual em duas dimenssões dos dados da nossa tabela, e fazer uma relação entre os três diferentes grupos de alimentos.



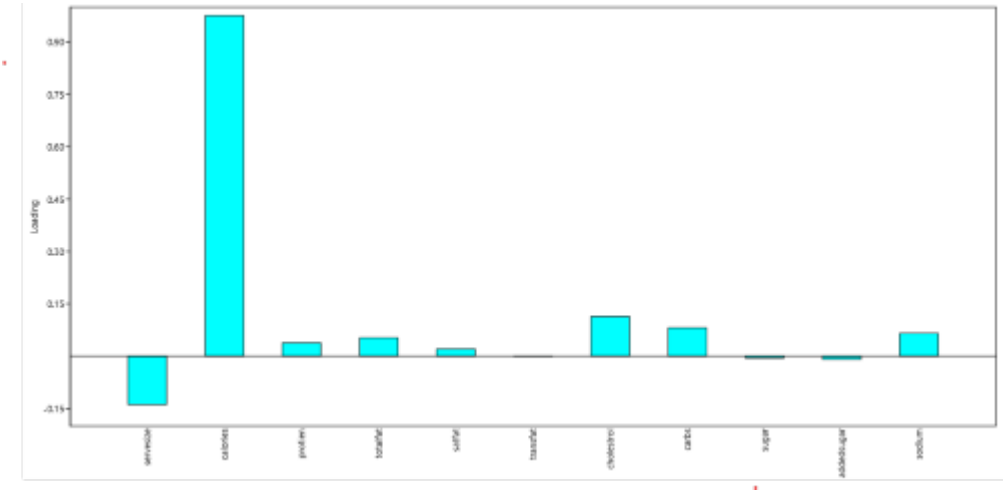
Clicando em Loadings plot você pode visualizar em forma de gráfico de barras qual coluna está puxando mais em cada componente principal. No nosso caso, o PC 1 está puxando muito mais as calorias, que tem uma relação inversamente proporcional ao tamanho da porção.



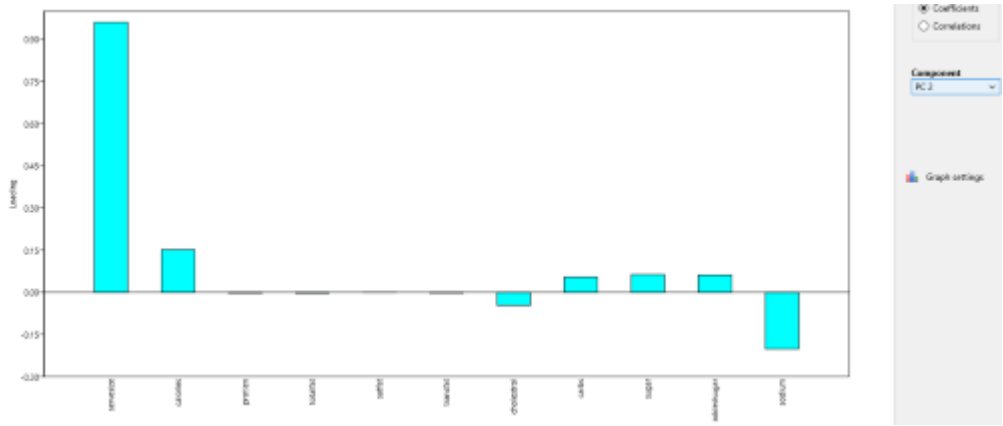
2.3 Análise do PCA do McDonalds

. Para fazer a análise do PCA primeiro iremos ver quais colunas estão puxando mais nos componentes principais que vamos selecionar, nesse primeiro caso iremos trabalhar com PC1 e o PC2.

Loadings Plot do PC1:

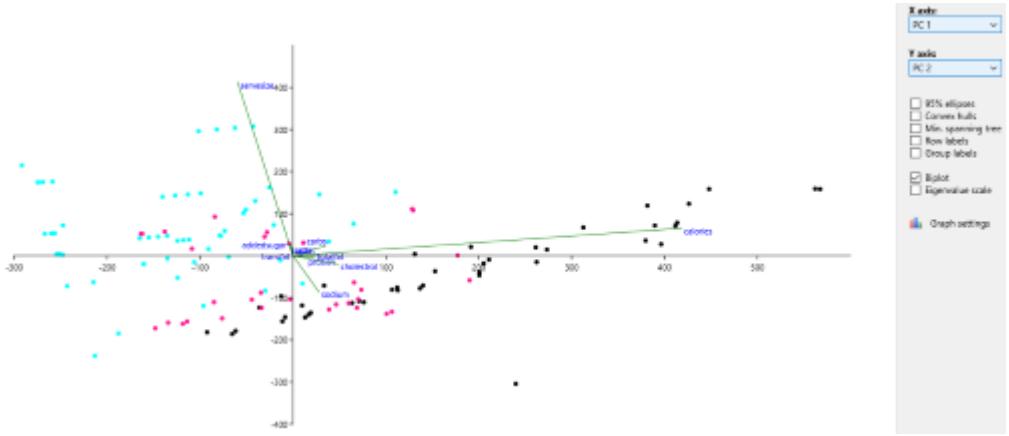


Loadings Plot do PC2:

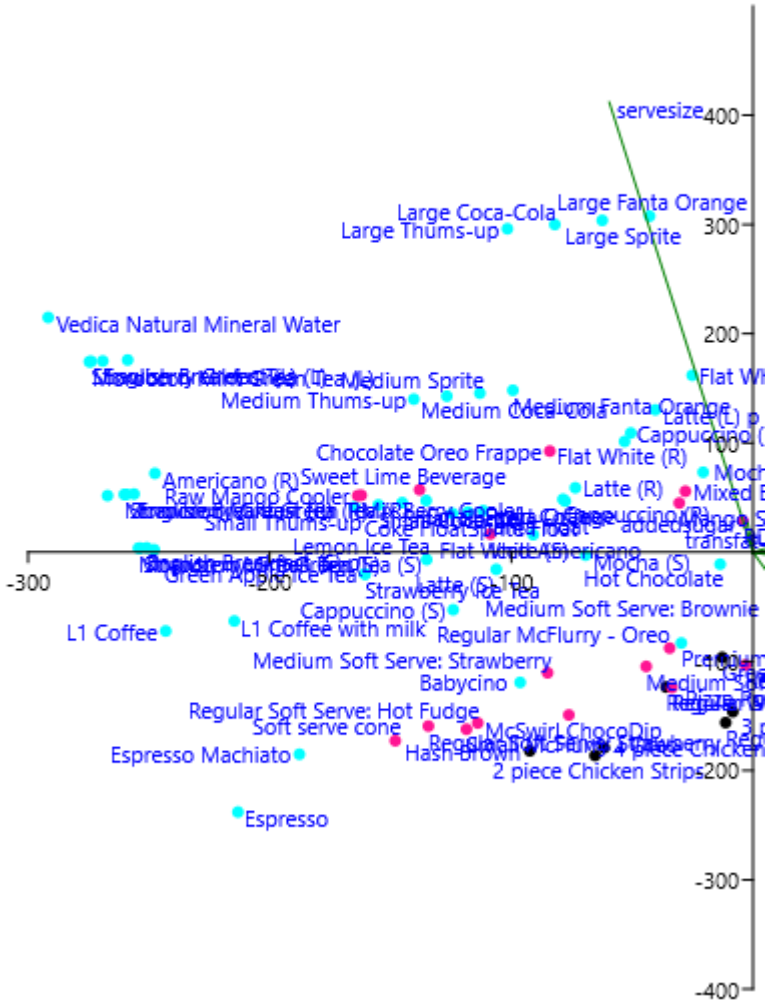


No PC1 podemos ver uma relação inversamente proporcional entre os dados das calorias com a porção que é servido, isso se dá pelo fato dos alimentos sólidos terem um número maior de calorias devido a sua quantidade de micro e macro nutrientes, já os líquidos tem um quantidade menor de calorias, mas são servidos em grandes porções na rede de fast food McDonalds.

No PC2 podemos algo parecido, mas no caso dele a relação inversamente proporcional é entre o tamanho da porção e a quantidade de sódio, as bebidas possuem uma quantidade de sódio muito baixa se comparado com o tamanho da porção, já os alimentos salgados possuem uma quantidade elevada pelo tamanho da porção.



Podemos ver que os Líquidos, os azuis claros, estão posicionados mais a esquerda do eixo x, na parte negativa, indo na direão oposta das calorias. Já os lanches salgados, pontos pretos, fazem o caminho contrário, e vão na direção oposta ao tamanho da porção, pois possuem uma quantidade calórica muito alta em comparação com o tamanho de sua porção. No meio disso temos as sobremesas, de cor rosa, que possuem uma relação mais equilibrada entre a quantidade de calorias e o tamnho de sua porção, tendo a ficar mais do lado esquerdo, se aproximando do eixo (0,0) devido ao vetor de Sódio



Aqui vemos que a água vedica é o ponto mais distante do vetor das calorias, o que faz sentido já que a água possui suas calorias zeradas, mas o tamanho da sua porção não é tão granfe quanto a dos refrigerantes.