Cel ćwiczenia: przeprowadzenie analizy głównych składowych (Principal component analysis) z użyciem zbioru danych „Cereals” zawierającego 30 rodzajów płatków śniadaniowych wraz z informacjami o kaloriach, ilości białka, tłuszczu, błonnika, węglowodanów, cukrów (w gramach) i sodu (w miligramach); (porcja: 30g.) Dane zostały ówcześnie przygotowane: usunięcie danych odstających, transformacja (o ile tego wymagały).

Dane zostały poddane autoskalowaniu przed HCA (nie ponawiam tego kroku przy PCA).

Celem PCA jest taki obrót układu współrzędnych, aby maksymalizować w pierwszej kolejności wariancję pierwszej współrzędnej, następnie wariancję drugiej współrzędnej itd..

Pierwszym krokiem w PCA jest wykonanie macierzy korelacji, niezbędnej do obliczenia wartości własnych i wektorów własnych macierzy w kolejnym kroku.

	calories	protein	fat	sodium*	fiber	carbo*	sugars
calories	1.000000	-0.235053	0.519579	0.064133	-0.167451	0.233747	0.343822
protein	-0.235053	1.000000	0.004077	-0.187741	0.750568	-0.082729	-0.523169
fat	0.519579	0.004077	1.000000	-0.152857	0.102935	0.528355	0.136245
sodium*	0.064133	-0.187741	-0.152857	1.000000	-0.112986	-0.409305	-0.211934
fiber	-0.167451	0.750568	0.102935	-0.112986	1.000000	0.124865	-0.260098
carbo*	0.233747	-0.082729	0.528355	-0.409305	0.124865	1.000000	0.743915
sugars	0.343822	-0.523169	0.136245	-0.211934	-0.260098	0.743915	1.000000

Tabela 1. Macierz korelacji dla cech (zbiór danych: "Cereals") m x m

Dzięki macierzy korelacji obliczyłam wektory własne i odpowiadające im wartości własne.

Posortowane malejąco (wraz ze spadkiem wyjaśnionej wariancji) wartości własne wraz z ich wektorami własnymi.

2.519 [0.377 0.03 -0.589 -0.181 -0.664 -0.22 -0.021]

1.928 [-0.325 0.523 -0.074 -0.003 -0.227 0.237 0.673]

1.189 [0.36   0.348 -0.499 -0.261 0.56   0.358 -0.094]

0.684 [-0.189 -0.37 -0.5   0.687 0.239 -0.002 0.237]

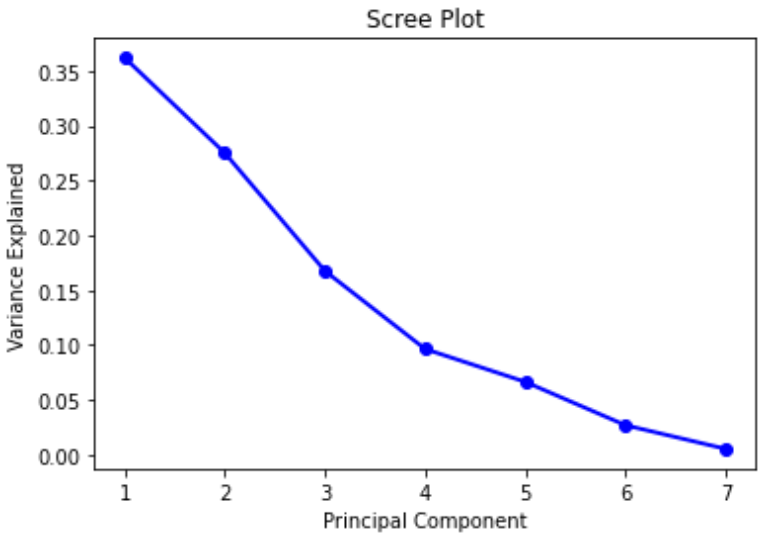
0.466 [-0.194 0.572 -0.067 0.431 -0.158 -0.004 -0.626]

0.187 [0.538 -0.019 0.316 0.395 -0.227 0.618 0.031]

0.038 [0.51   0.373 0.21   0.293 0.244 -0.621 0.298]

Kolejnym krokiem będzie wybór optymalnej liczby głównych składowych. Aby ustalić ich ilość użyłam dwóch metod:

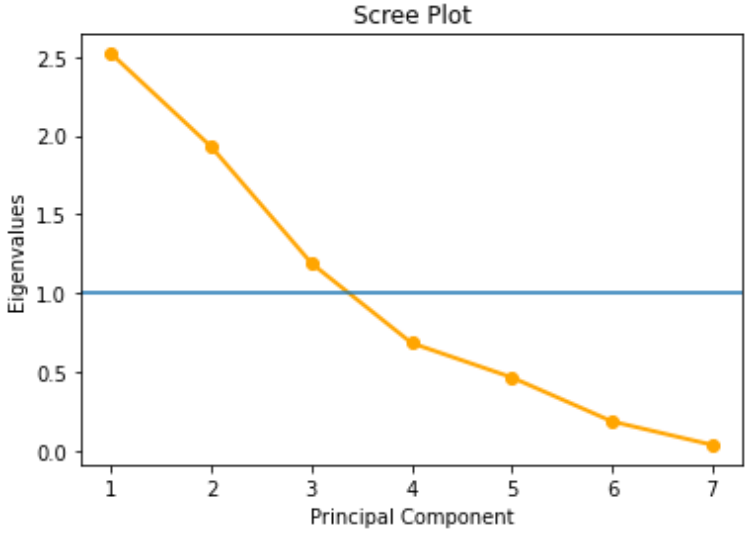
- kryterium minimalnego zasobu zmienności – przyjęłam minimalny procent zmienności jaką chciałabym określić i na tej podstawie dobrałam liczbę głównych składowych (k).



Minimalny procent: 80%

Po zsumowaniu wyjaśnianej zmienności przez pierwsze trzy główne składowe uzyskałam ustalone przez siebie minimum. 3 pierwsze główne składowe wyjaśniają 80,4 % ogólnej zmienności.

- kryterium Keiser'a - jako istotne główne składowe wybieramy tylko te, których wartości własne $\geqslant 1$.

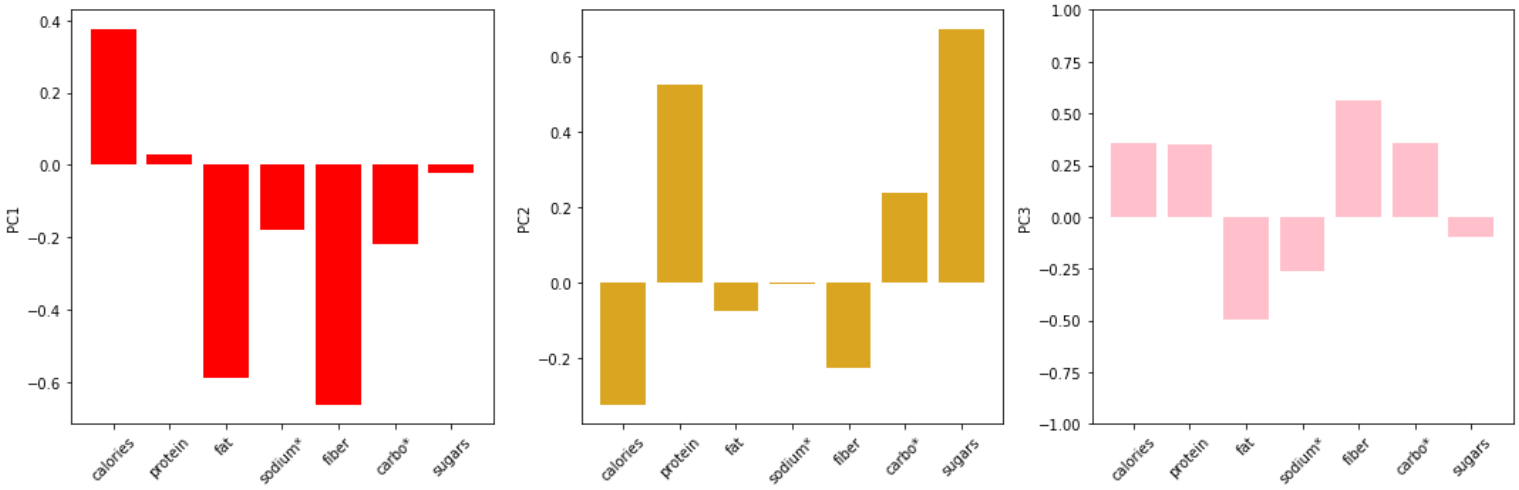


Przy zastosowaniu kryterium Keiser'a również wybieram jako istotne 3 pierwsze główne składowe.

Następnie obliczyłam ładunki czynnikowe, czyli współczynniki korelacji pomiędzy daną zmienną, a składowymi. Stworzyłam macierz ładunków oraz wykresy słupkowe ładunków czynnikowych.

	PC1	PC2	PC3
calories	0.377	-0.325	0.360
protein	0.030	0.523	0.348
fat	-0.589	-0.074	-0.499
sodium*	-0.181	-0.003	-0.261
fiber	-0.664	-0.227	0.560
carbo*	-0.220	0.237	0.358
sugars	-0.021	0.673	-0.094

Tabela 2. Macierz ładunków k x m



Rysunek 1. Wykresy ładunków czynnikowych dla poszczególnych składowych.

Co prawda żaden z ładunków nie spełnia kryterium Malinowskiego – minimalna wartość (0.7) po pokonaniu której zachodzi korelacja pomiędzy zmienną ukrytą, a zmienną oryginalną, można jednak wyciągnąć następujące wnioski:

Pierwszy wektor własny (PC1) ma wysokie ładunki dodatnie przy zmiennej „calories” oraz wysokie ujemne ładunki przy zmiennych „fat” i fiber” . Kalorie i tłuszcz wykazują dodatnią korelację. Tłuszcz i błonnik wykazują ujemną korelację, kalorie i błonnik wykazują dodatnią korelację, zmienne te mają największy wpływ na PC1.

Drugi wektor własny (PC2) ma wysokie ładunki dodatnie przy zmiennej „protein”, „sugars”. Białka i cukry wykazują ujemną korelację. Zmienne te mają największy wpływ na PC2.

Trzeci wektor własny (PC3) ma mały ładunek dodatni przy zmiennych: „calories”, „protein”, „fiber” i „carbo” oraz mały ujemny ładunek przy zmiennych „fat”, „sodium” i „sugars”. Interpretacja tej zmiennej nie jest oczywista.

Biplots

„Biplots” – nakłada na siebie wykres punktów utworzony wg. nowych współrzędnych i wykres ładunków.

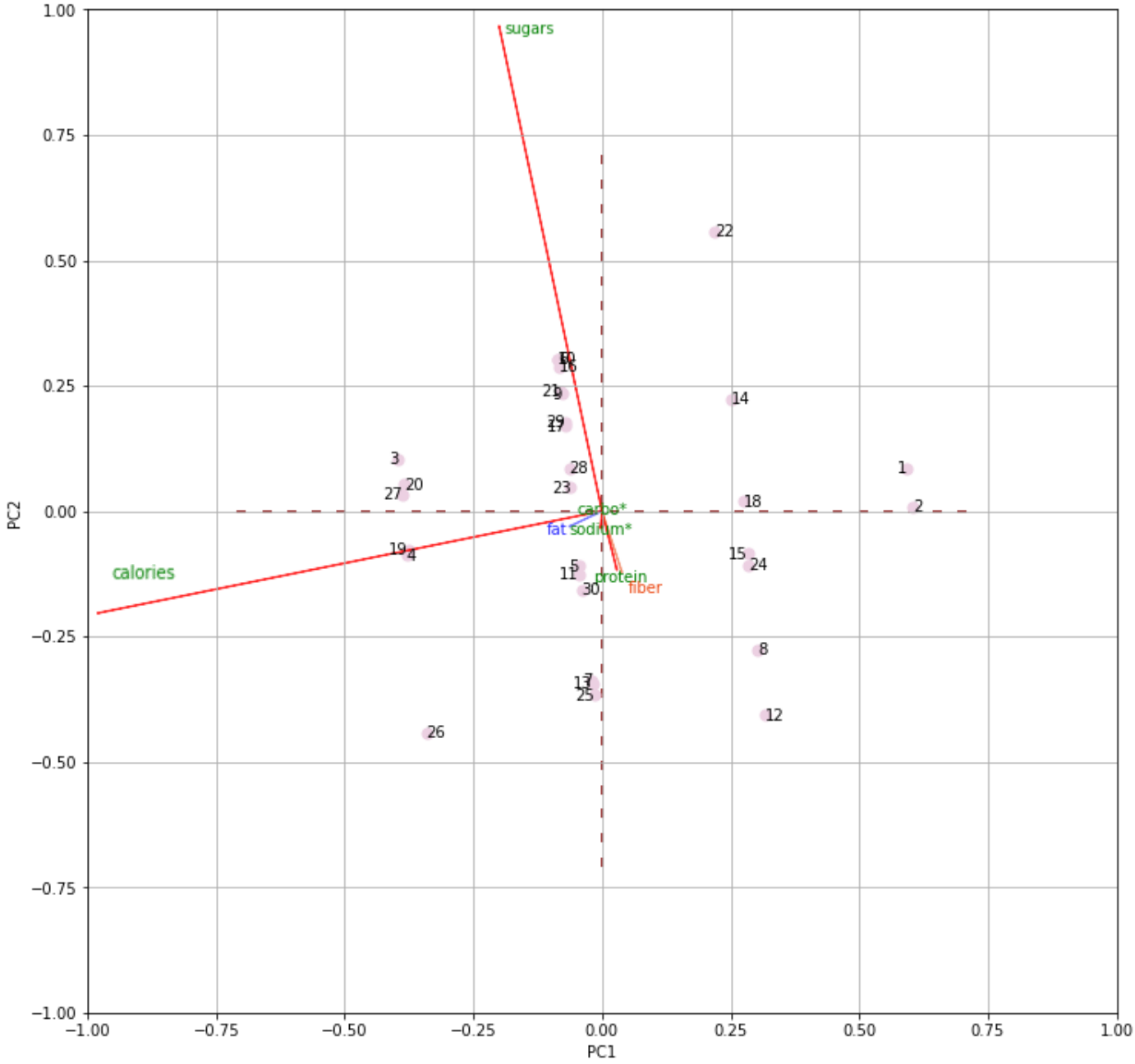
Dostarcza więcej informacji niż dwa wykresy interpretowane samodzielnie.

Oznaczenia płatków na wykresach:

1. Bran Chex
2. Bran Flakes
3. Cap'n'Crunch
4. Cinnamon Toast Crunch
5. Clusters
6. Cocoa Puffs
7. Corn Chex
8. Corn Flakes
9. Corn Pops
10. Count Chocula
11. Cracklin' Oat Bran
12. Cream of Wheat (Quick)
13. Crispix
14. Crispy Wheat & Raisins
15. Double Chex
16. Froot Loops
17. Frosted Flakes
18. Frosted Mini-Wheats
19. Fruit & Fibre Dates; Walnuts; and Oats
20. Fruitful Bran
21. Fruity Pebbles
22. Golden Crisp
23. Golden Grahams
24. Grape Nuts Flakes
25. Grape-Nuts
26. Great Grains Pecan
27. Honey Graham Ohs
28. Honey Nut Cheerios
29. Honey-comb
30. Just Right Crunchy Nuggets

Kolory strzałek i nazwy zmiennych na wykresach są różne, aby biplot był bardziej czytelny.

PC1 i PC2



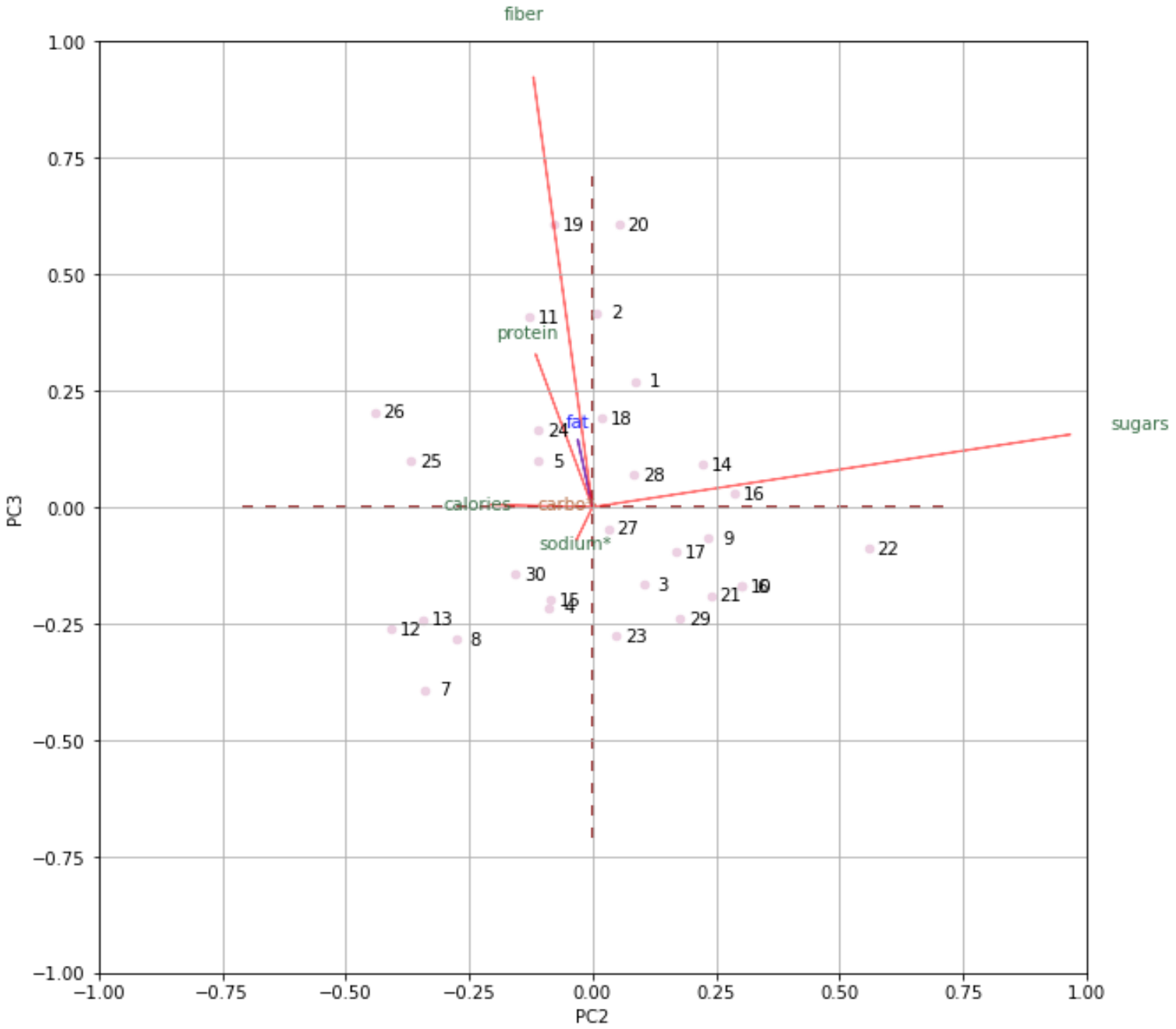
Wartości PC1 dla zmiennych wahają się od -0.30 do 0.65, natomiast PC2 w przedziale -0.5 do 0.6. Na wykresie widzimy wyodrębnione 4 skupiska. Różnica w wartości PC1 obiektów w każdym ze skupisk jest niewielka, o wiele większa jest różnica w wartościach PC2.

Pierwsze skupisko (wysunięte najbardziej na lewo) utworzone z numerów 3,4,19,20,26,27 to płatki o najwyższej ilości kalorii, zwracając uwagę na wykres ładunków PC1 możemy stwierdzić, że płatki te mają także dość dużą zawartość tłuszczu. (kalorie skorelowane są dodatnio z tłuszczem).

Skupisko najbardziej wysunięte na prawo złożóne jedynie z dwóch rodzajów płatków: 1 i 2 ma najniższą ilość kalorii. Łatwo więc o interpretację – im obiekt położony bardziej na prawo (większa wartość PC1) – tym ilość kalorii jest większa.

Płatki 10,16,22 mają największą zawartość cukrów, widać to tu, jak i na wykresie PC2 do PC3.

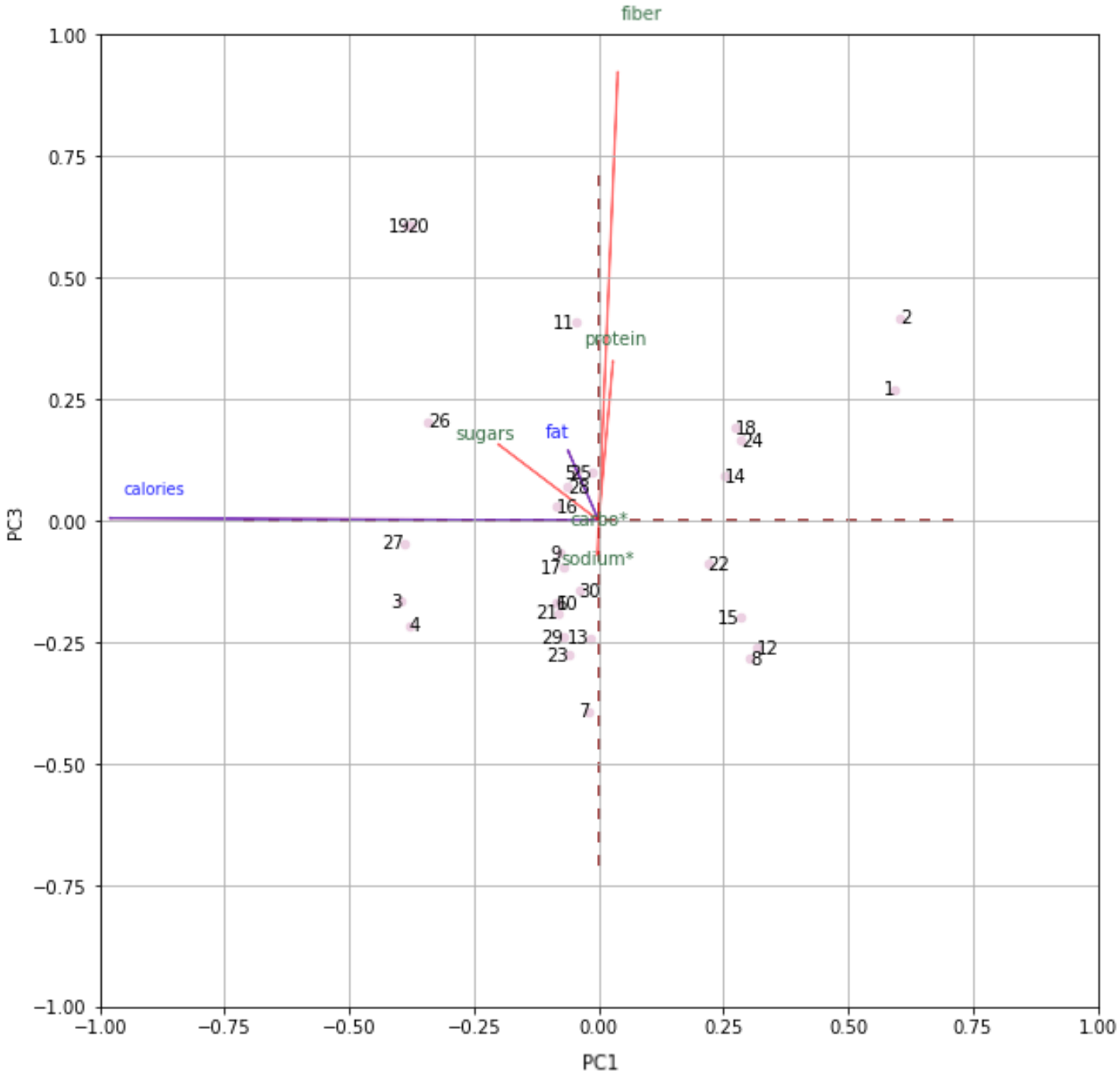
PC2 i PC3



Zgodnie z kierunkiem strzałki wskazującej na zmienną sugars - im płatki bardziej wysunięte na prawo tym ilość cukrów jest większa.

Długa strzałka fiber skierowana ku górze wskazuje na dużą zawartość błonnika (duża wartość PC3) – płatki 2, 11, 19, 20 mają największą ilosć błonnika ze wszystkich płatków w zbiorze.

PC1 i PC3



Płatki o najniższej wartości PC1 to płatki z najwyższą ilością kalorii (120 kcal). Płatki o najwyższej wartości PC1 to płatki o najniższej ilości kalorii (90 kcal). Im wyższa wartość PC3 tym wyższa wartość błonnika.

Skupisko stworzone z płatków 22, 15, 12 i 8 to płatki o zerowej zawartości tłuszczu i tym samym niskiej zawartości kalorii.

Płatki numer 11 – stosunkowo dużo tłuszczu jak i białka równocześnie.

Podobnie jak na wykresie PC1 i PC2 również i tu formują się 4 skupiska z obiektami o podobnej wartości PC1, dzieje się tak ze względu na obecność wektora PC1, na który duży wpływ mają kalorie i tłuszcz w płatkach.