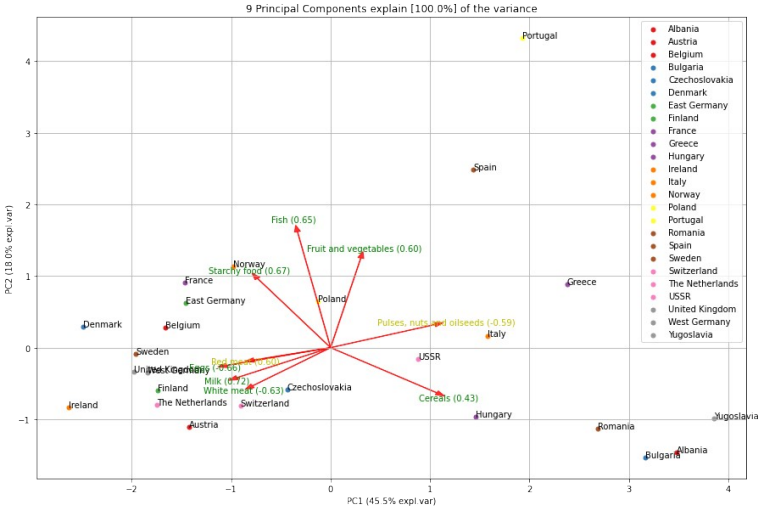
Principale componenten analyse
Oplossingen

1 Protein consumption

a) let op de separator en hoe kommagetallen genoteerd zijn!
b) je kan dit doen door protein.index aan te passen
c) Country bevat enkel nominale gegevens en Total is vrij onzinnig omdat het gewoon de som van de andere kolommen is. Deze kolommen mogen dus weg.
d) er zijn heel wat duidelijke correlaties tussen de kolommen. Ideaal voor PCA dus.
e) we hebben minstens 5 componenten nodig om 90% te halen
f) de eerste component is een lineaire combinatie van de oorspronkelijke kolommen. De gewichten zijn allemaal wat van dezelfde grootorde. Er steekt geen kolom specifiek uit.
g)



h) je kan dit doen door de juiste kolommen te selecteren of door de PCA opnieuw uit te voeren met maar 3 componenten.
i) Als je maar 3 componenten selecteert en hier 6 clusters zoekt, dan vind je:

cluster 0 : ['Albania']
cluster 1 : ['Austria', 'Belgium', 'Czechoslovakia', 'Denmark', 'East Germany', 'Finland', 'France', 'Ireland', 'The Netherlands', 'Norway', 'Poland', 'Sweden', 'Switzerland', 'United Kingdom', 'West Germany']
cluster 2 : ['Bulgaria', 'Greece', 'Italy', 'Romania', 'USSR', 'Yugoslavia']
cluster 3 : ['Hungary']
cluster 4 : ['Portugal']
cluster 5 : ['Spain']

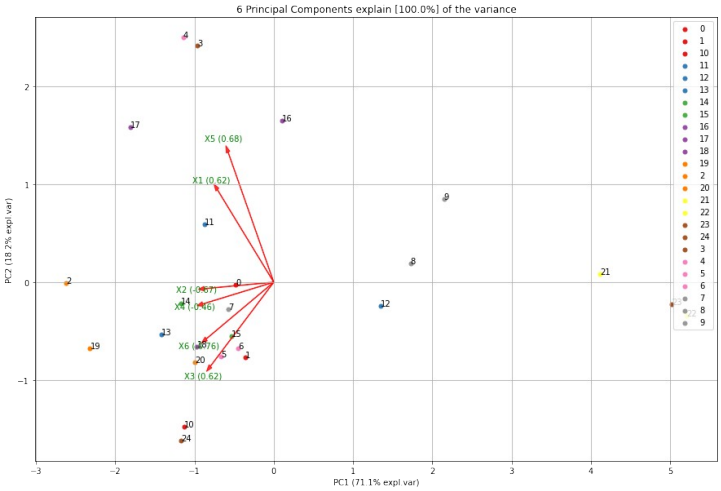
Als je de oorspronkelijke data gebruikt, vind je volgende clusters:

cluster 0 : ['Albania']
cluster 1 : ['Austria', 'Belgium', 'Czechoslovakia', 'Denmark', 'East Germany', 'Finland', 'France', 'Ireland', 'The Netherlands', 'Norway', 'Poland', 'Sweden', 'Switzerland', 'United Kingdom', 'USSR', 'West Germany']
cluster 2 : ['Bulgaria', 'Hungary', 'Romania', 'Yugoslavia']
cluster 3 : ['Greece', 'Italy']
cluster 4 : ['Portugal']
cluster 5 : ['Spain']

Er is dus veel gelijkenis tussen de gevonden clusters. Met slechts 3 componenten kan je de data dus al vrij goed klassificeren. Deze drie componenten verklaren 75,6% van de variantie.

2 Goblets

a) let weer op de separator
b) de kolom Goblet bevat volgnummers en is dus nominaal. Deze mag weg
c) we hebben minstens 3 componenten nodig
d)

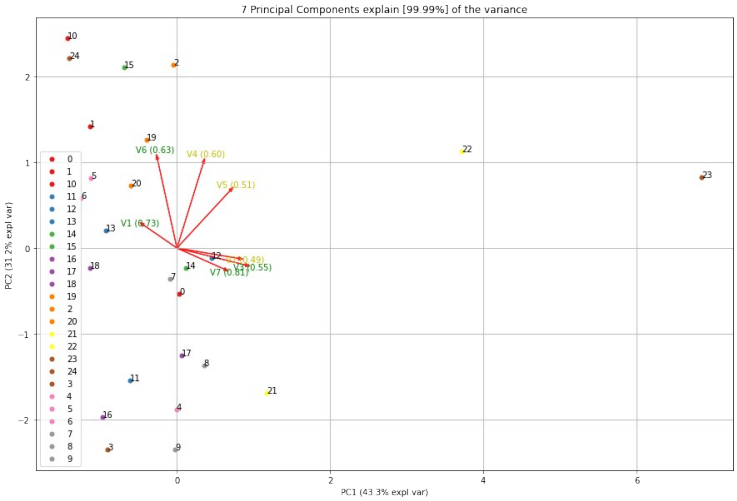


e) De nieuwe tabel ziet er zo uit:

	V1	V2	V3	V4	V5	V6	V7
0	1.615385	1.500000	0.928571	3.000000	1.857143	2.000000	2.875000
1	1.000000	0.736842	0.736842	2.800000	2.800000	3.800000	2.666667
2	1.210526	1.150000	0.950000	3.833333	3.166667	3.333333	2.000000
3	1.058824	1.125000	1.062500	1.636364	1.545455	1.454545	2.000000
4	1.052632	1.250000	1.187500	2.000000	1.900000	1.600000	2.285714
5	1.666667	1.176471	0.705882	3.333333	2.000000	2.833333	2.666667

...

f) Nu hebben we 4 componenten nodig, maar dan hebben we wel ineens 99,6% van de verklaarde variantie
g)



h) Bij de oorspronkelijke tabel waren de correlaties tussen de kolommen (gemiddeld 0,700) hoger dan bij de tabel verhoudingen (gemiddeld 0,297). Dat uit zich in de verklaarde variantie van de eerste component die 43,4% was bij verhoudingen en 71,2% bij de oorspronkelijke tabel. Hoe hoger de correlaties tussen de kolommen, hoe beter PCA in staat is om het aantal kolommen te verminderen.

3. CPU’s

a) de kolom ‘name’ moet verwijderd worden

b) er zijn heel wat correlaties te vinden in de tabel (de gemiddelde correlatie is 0,475). Prima voor PCA dus

c) we hebben 5 componenten nodig

d) hier steekt ook geen enkele kolom uit. De coëfficiënten liggen allemaal tussen 0.199 en 0.423

e)