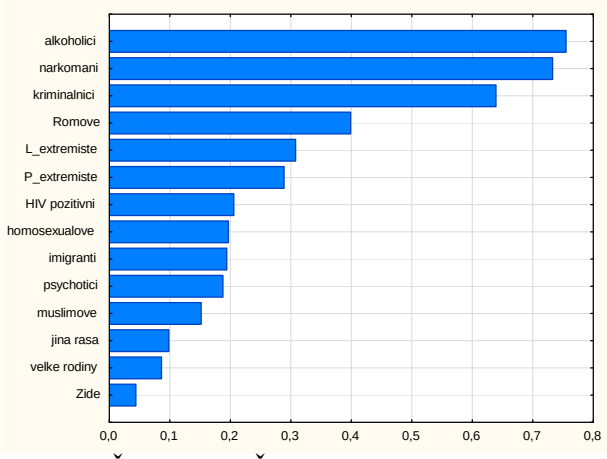
Cvičení č. 12.: Aplikace shlukové analýzy

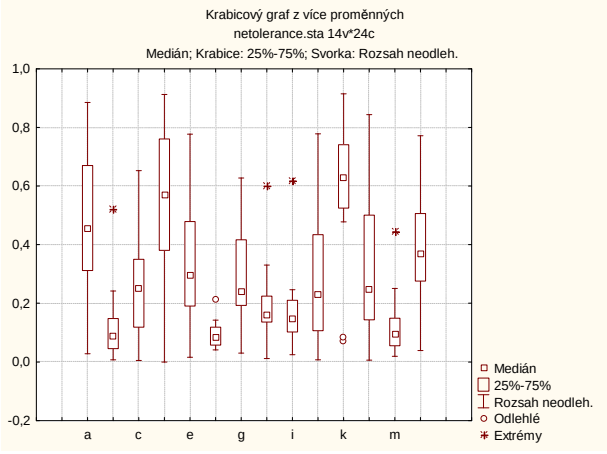
Článek Ladislava Rabušice Koho Češi nechtějí? (uveřejněn ve Sborníku prací FSS MU Sociální studia 5, 2000) se zabývá touto problematikou:
V roce 1999 proběhlo ve 24 evropských zemích sociologické šetření, v němž měli respondenti za úkol odpovědět na otázku „Můžete prosím z následujícího seznamu vybrat všechny ty, koho byste nechtěl(a) mít za sousedy?“ V seznamu byly tyto skupiny osob:
Kriminálníci, osoby jiné rasy, levicoví extrémisté, alkoholici, pravicoví extrémisté, početné rodiny, psychotici, muslimové, imigranti, HIV pozitivní, narkomani, homosexuálové, židé, Romové.
V datovém souboru netolerance.sta jsou zaznamenány relativní četnosti vybraných skupin osob.
V České republice se výzkumu, který proběhl v květnu 1999, zúčastnilo 1908 osob.

Úkol 1.: Zaměřte se na ČR. Vytvořte sloupkový diagram tohoto tvaru:



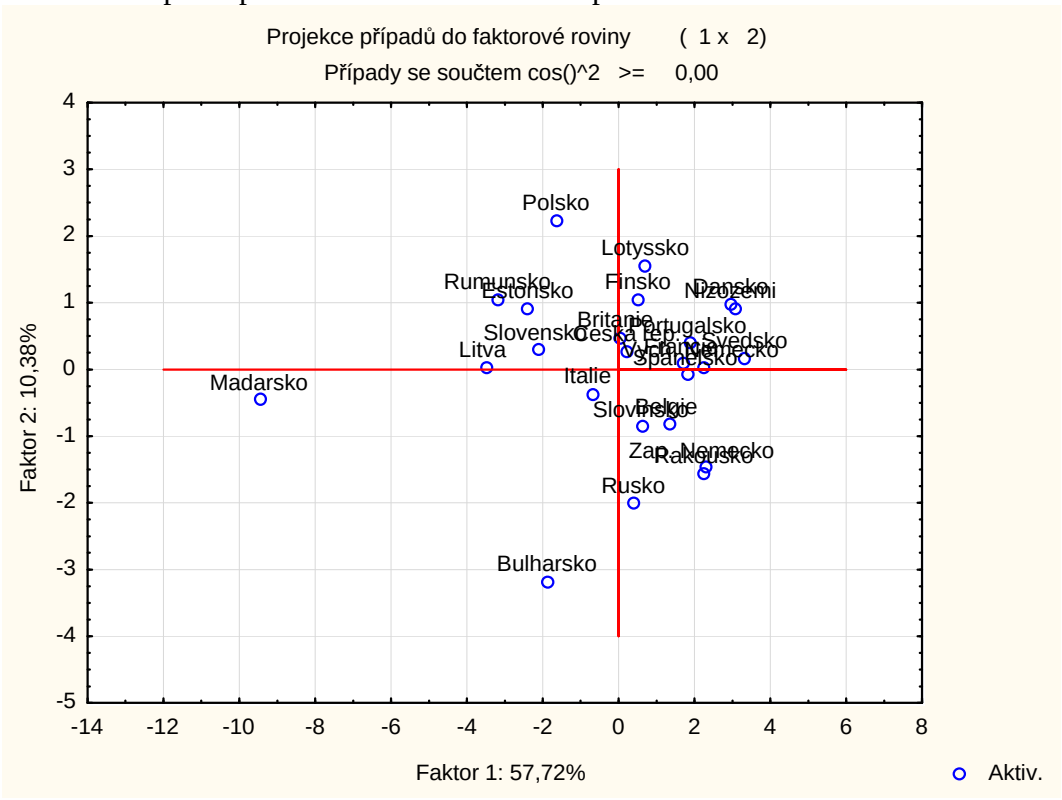
Návod: Řádek pro Českou republiku okopírujeme (se záhlavími) do nového datového souboru o 14 proměnných a jednom případu.
Soubor transponujeme: Data – Transponovat – Soubor.
Hodnoty proměnné Ceska rep. uspořádáme: Data – setřídit – Přidat prom. Ceska rep. – OK.
Nakreslíme sloupcový graf: Grafy – 2D grafy – Sloupcové/pruhové grafy – Proměnné Ceska rep. – O, Typ grafu Běžný, Orientace Horizontální – OK.

Úkol 2.: Do jednoho grafu nakreslete krabicové diagramy všech 14 proměnných.

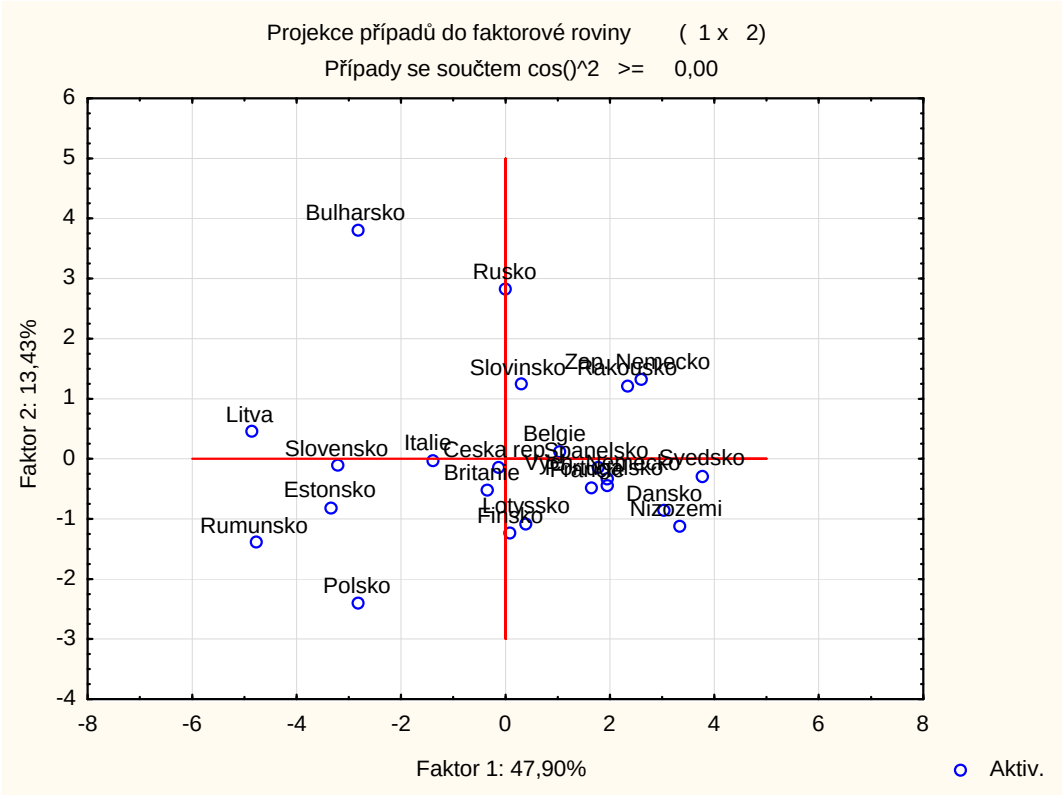


Vzhledem k velmi rozdílné variabilitě proměnných se jeví vhodnější pracovat se standardizovanými daty.

Úkol 3.: Na ploše prvních dvou hlavních komponent znázorněte rozmístění zemí.

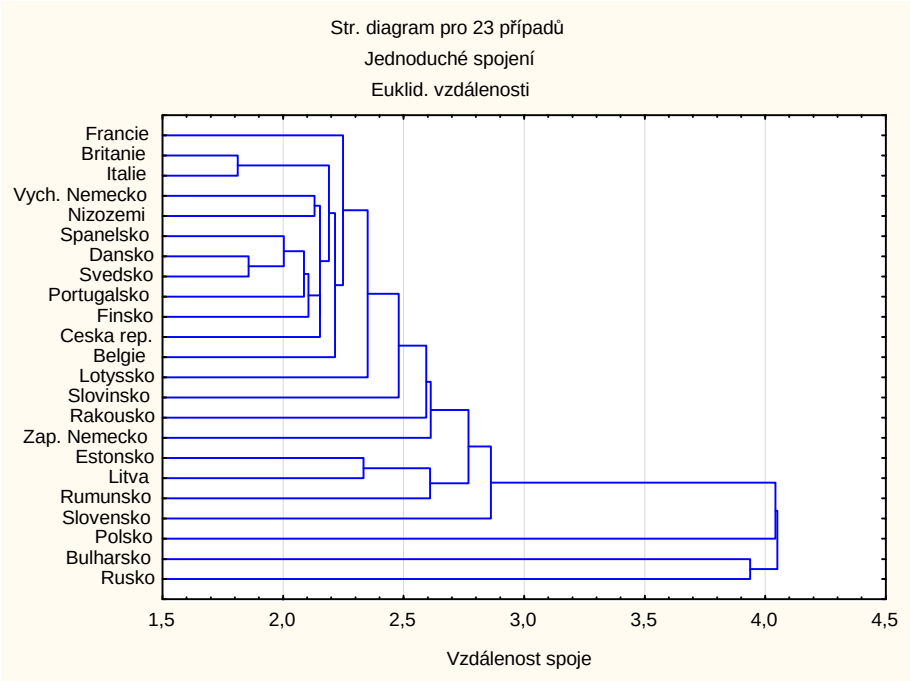


Maďarsko se jeví jako odlehlé pozorování. Z dalších analýz ho vyloučíme. Znovu provedeme metodu hlavních komponent a dostaneme toto rozmístění zemí:

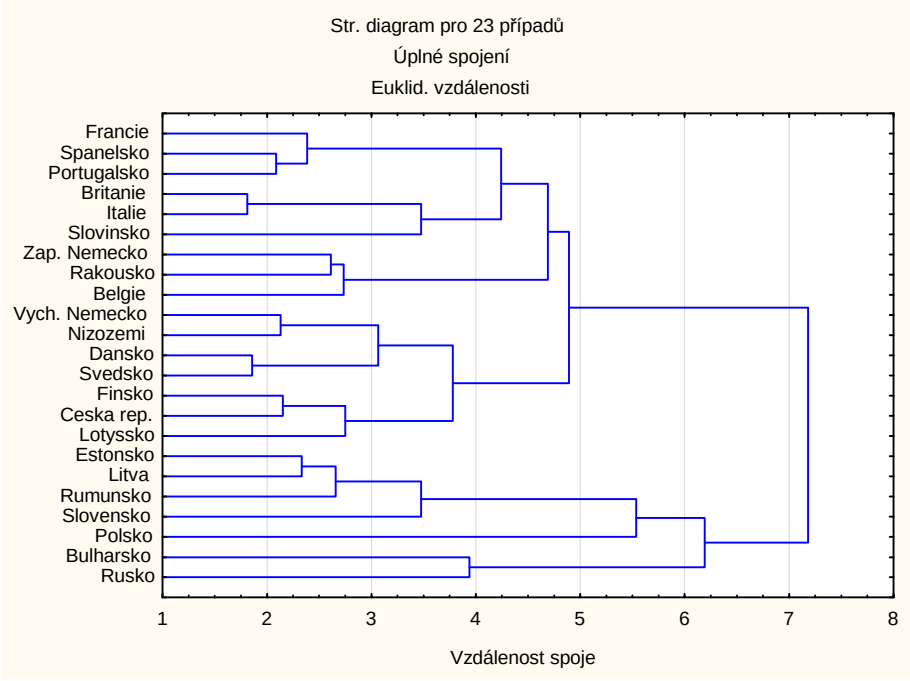


Úkol 4.: Použijte metodu nejbližšího souseda, nejvzdálenějšího souseda, metodu průměrné vazby a Wardovu metodu pro nalezení shluků zemí podobných z hlediska tolerance. Výsledky znázorněte pomocí dendrogramů.

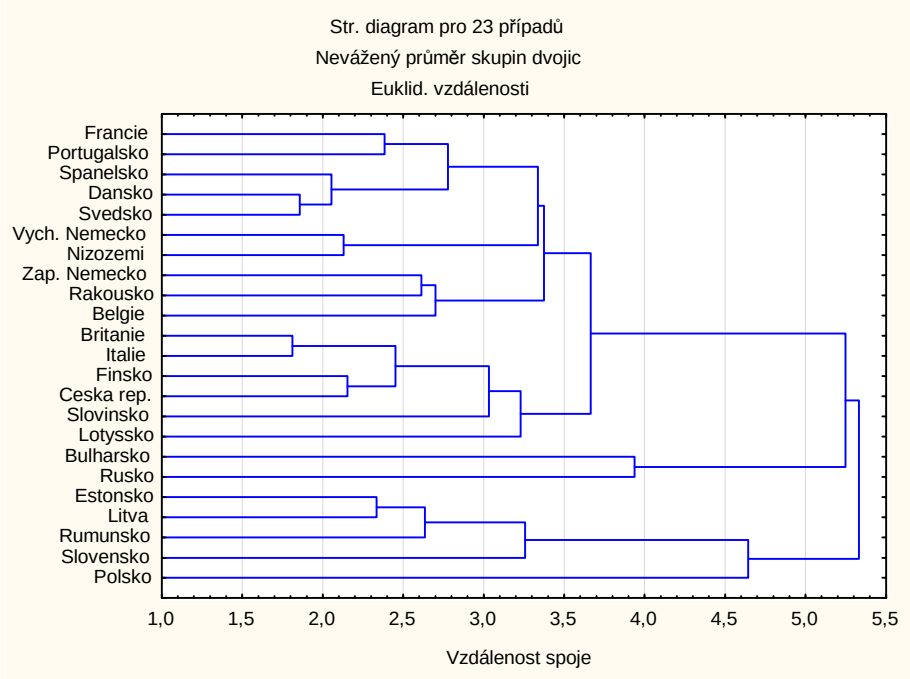
Dendrogram pro metodu nejbližšího souseda:



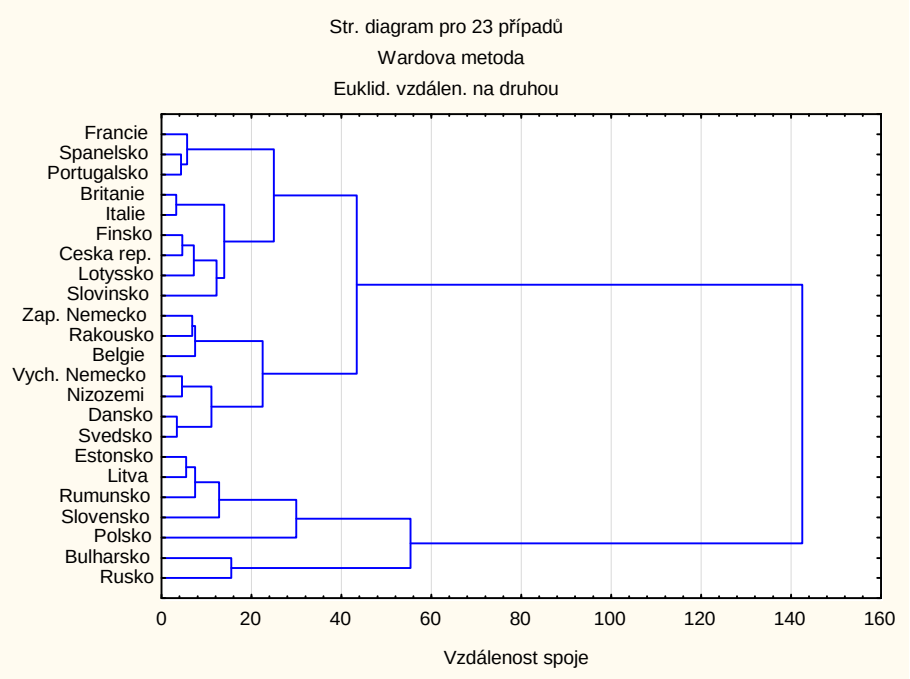
Dendrogram pro metodu nejvzdálenějšího souseda:



Dendrogram pro metodu průměrné vazby:

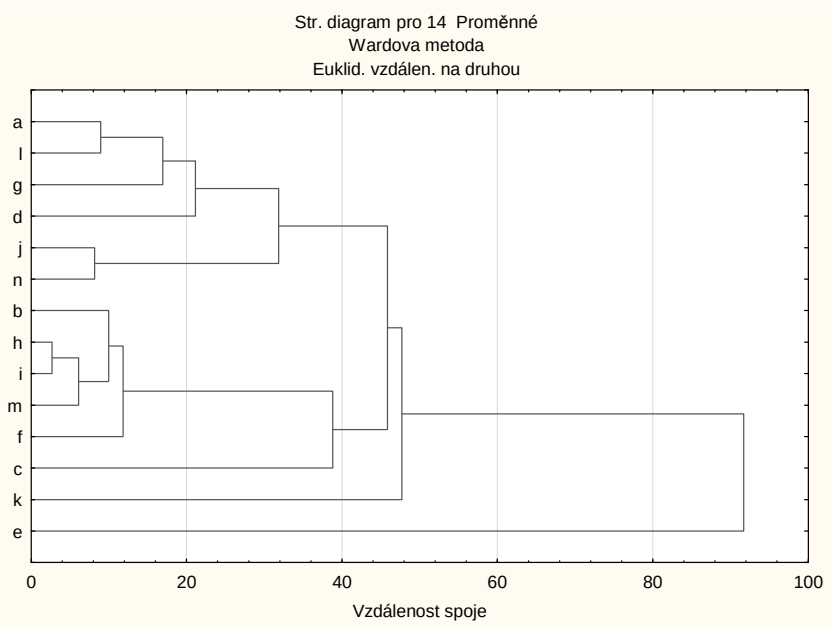


Dendrogram pro Wardovu metodu:



Úkol 5.: Pro Wardovu metodu určete 4 shluky navzájem si podobných zemí.
Shluk č. 1: Francie, Španělsko, Portugalsko, Velká Británie, Itálie, Finsko, ČR, Lotyšsko, Slovinsko
Shluk č. 2: Západní Německo, Rakousko, Belgie, Východní Německo, Nizozemí, Dánsko, Švédsko
Shluk č. 3: Estonsko, Litva, Rumunsko, Slovensko, Polsko
Shluk č. 4: Bulharsko, Rusko

Úkol 6.: Proveďte shlukovou analýzu pro proměnné.
Dendrogram pro Wardovu metodu:



Proměnné roztřídíme do pěti shluků.
Shluk č. 1: a (kriminálníci), l (homosexuálové), g (citově nestabilní lidé), d (alkoholici), j (lidé s AIDS), n (Romové)
Shluk č. 2: b (osoby jiné rasy), h (muslimové), i (imigranti), m (Židé), f (velké rodiny)
Shluk č. 3: c (levicoví extrémisté)
Shluk č. 4: k (narkomani)
Shluk č. 5: e (pravicoví extrémisté)

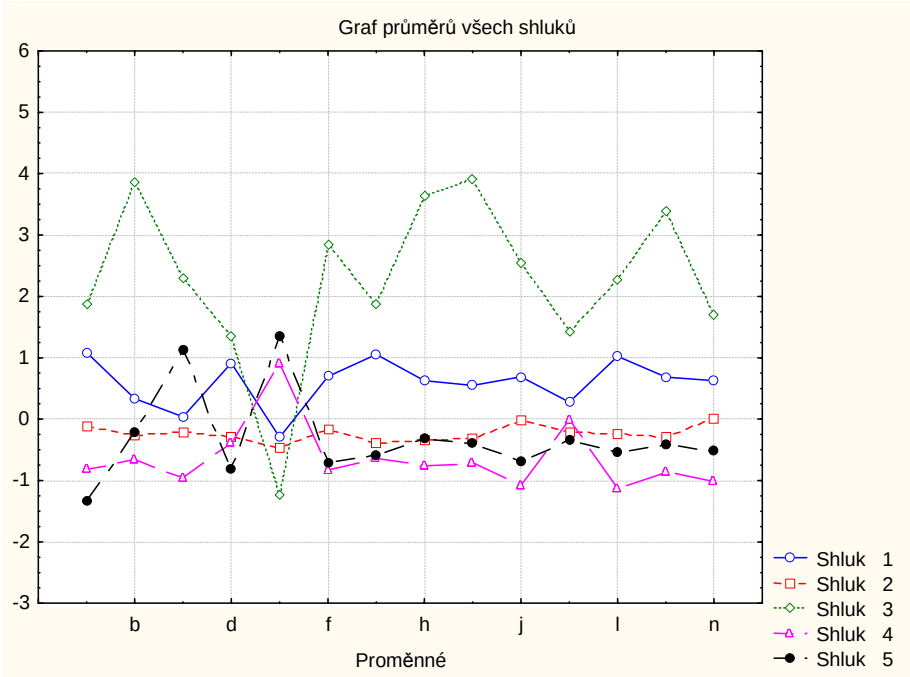
Úkol 7.: Použijte metodu k-průměrů k nalezení 5 shluků navzájem si podobných zemí a uložte skupinovou příslušnost do datového souboru. K určení významnosti jednotlivých proměnných proveďte analýzu rozptylu. Nakreslete graf průměrů všech 5 shluků a pokuste se o interpretaci.
Rozdíly oproti Wardově metodě: Bulharsko bylo zařazeno do shluku č. 3, Rusko do shluku č. 1 a shluk č. 2 se rozpadl na dva.

Výsledek analýzy rozptylu:

Proměnná	Analýza rozptylu (netolerance.sta)					
	Mezisk. SČ	sv	Vnitřní SČ	sv	F	význam. p
a	18,65740	4	4,34260	19	20,40772	0,000001
b	18,19146	4	4,80854	19	17,96998	0,000003
c	13,24487	4	9,75513	19	6,44924	0,001872
d	10,32416	4	12,67584	19	3,86876	0,018332
e	12,98238	4	10,01762	19	6,15579	0,002368
f	15,51020	4	7,48980	19	9,83650	0,000174
g	14,40255	4	8,59745	19	7,95726	0,000605
h	19,38696	4	3,61304	19	25,48770	0,000000
i	20,57442	4	2,42558	19	40,29076	0,000000
j	15,46437	4	7,53563	19	9,74780	0,000184
k	3,25872	4	19,74128	19	0,78409	0,549512
l	17,93810	4	5,06190	19	16,83280	0,000005
m	18,58430	4	4,41570	19	19,99126	0,000001
n	10,05713	4	12,94287	19	3,69094	0,021879

Na hladině významnosti 0,05 není významná pouze proměnná k (narkomani). Podle hodnot statistiky F lze soudit, že na zařazování zemí do shluků se nejvíce podílí proměnná i (imigranti), dále h (muslimové), a (kriminálníci), m (Židé), b (osoby jiné rasy) a l(homosexuálové).

Graf průměrů:



Vidíme, že velmi výrazně se odlišuje shluk č. 3 (Maďarsko), které s výjimkou proměnné e (pravicoví extrémisté) vykazuje vyšší míru netolerance než je průměr shluků ostatních zemí. Naopak, země, které patří do shluku č. 4 (Vých. Německo, Nizozemí, Dánsko, Švédsko) mají – až na několik výjimek – vyšší míru tolerance než ostatní země.

Příklad k samostatnému řešení:

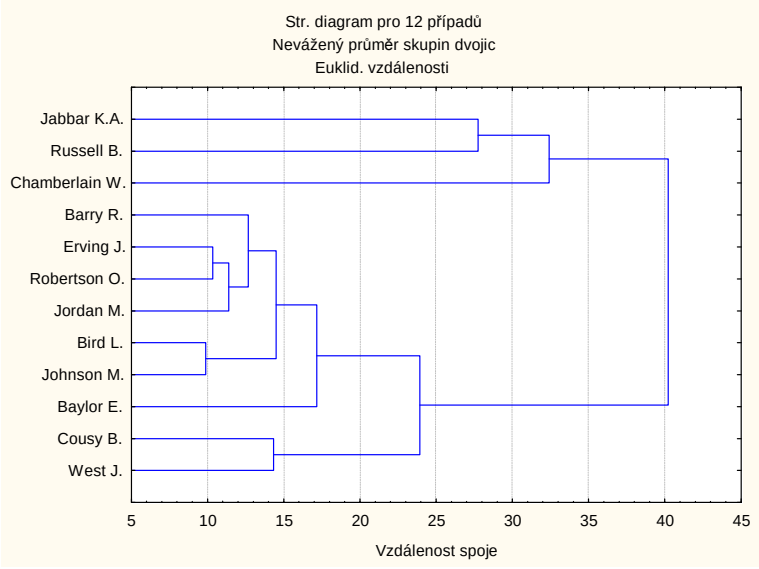
(Příklad je převzat z knihy M. Meloun, J. Militký, M. Hill: Počítačová analýza vícerozměrných dat. Academia Praha 2005)

U 12 velmi slavných amerických hráčů košíkové byly v sezóně 1989 zjištěny hodnoty osmi proměnných.
Výška – výška hráče v cm
Hmotnost – hmotnost hráče v kg
FgPct – první antropometrická charakteristika
FtPct – druhá antropometrická charakteristika
Body – průměrný počet dosažených bodů
Doskoky - průměrný počet doskoků
Asistence – průměrný počet asistencí
Fauly – průměrný počet faulů
Data jsou uložena v souboru hraci.sta.

	1 Jméno hráče	2 Vyska	3 Hmotnost	4 Fgpct	5 Ftpct	6 Body	7 Doskoky	8 Asistence	9 Fauly
1	Jabbar K.A.	218,6	105,0	55,9	72,1	24,6	11,2	3,6	3
2	Barry R.	200,8	93,6	44,9	90,0	23,2	6,7	4,9	3
3	Baylor E.	195,7	102,7	43,1	78,0	27,4	13,5	4,3	3,1
4	Bird L.	205,9	100,4	50,3	88,0	25,0	10,2	6,1	2,7
5	Chamberlain W.	216,0	125,5	54,0	51,1	30,1	22,9	4,4	2
6	Cousy B.	184,3	79,9	37,5	80,3	18,4	5,2	7,5	2,4
7	Erving J.	199,5	91,3	50,6	77,8	24,2	8,5	4,2	2,8
8	Johnson M.	205,9	98,1	53,0	83,4	19,5	7,4	11,2	2,4
9	Jordan M.	198,3	89,0	51,3	84,8	32,6	6,2	5,9	3,1
10	Robertson O.	195,7	95,8	48,5	83,8	25,7	7,5	9,5	2,8
11	Russell B.	207,1	100,4	44,0	56,1	15,1	22,6	4,3	2,7
12	West J.	189,4	82,2	47,4	81,4	27,0	5,8	6,7	2,6

Metodou průměrné vazby s euklidovskými vzdálenostmi najděte 3 skupiny hráčů podobných vlastností. Výsledek ověřte metodou k-průměrů. Zjistěte, které proměnné se nejvíce podílejí na zařazování hráčů do shluků

Výsledky
Dendrogram:



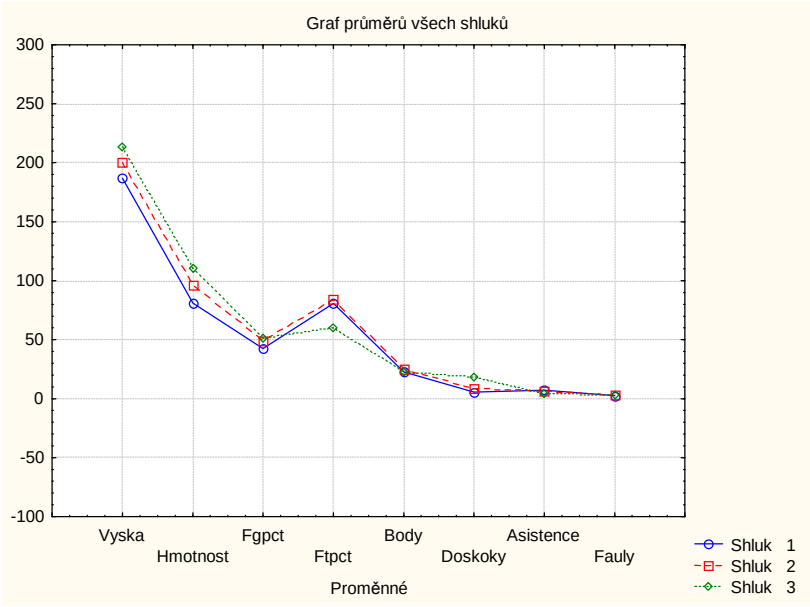
Rozdělení hráčů do 3 shluků metodou k-průměrů:

	Členy shluku číslo 1 (hraci.sta) a vzdálenosti od přislušného středu shluku Shluk obsahuje 2 příp.
	Vzdálen.
Cousy B.	2,532710
West J.	2,532710

	Členy shluku číslo 2 (hraci.sta) a vzdálenosti od přislušného středu shluku Shluk obsahuje 7 příp.
	Vzdálen.
Barry R.	2,995406
Baylor E.	4,557197
Bird L.	3,089724
Erving J.	2,877904
Johnson M.	3,738602
Jordan M.	3,819170
Robertson O.	1,951357

	Členy shluku číslo 3 (hraci.sta) a vzdálenosti od přislušného středu shluku Shluk obsahuje 3 příp.
	Vzdálen.
Jabbar K.A.	5,967011
Chamberlain W.	6,905056
Russell B.	6,030139

Graf průměrů tří shluků:



Tabulka ANOVA:

Proměnná	Analýza rozptylu (hraci.sta)					
	Mezisk. SČ	sv	Vnitřní SČ	sv	F	význam. p
Vyska	905,409	2	194,4173	9	20,95668	0,000411
Hmotnost	1051,052	2	505,9978	9	9,34734	0,006358
Fgpct	97,229	2	207,9136	9	2,10439	0,177914
Ftpct	1232,846	2	368,0602	9	15,07310	0,001340
Body	16,239	2	249,3210	9	0,29310	0,752805
Doskoky	287,475	2	127,7543	9	10,12598	0,004970
Asistence	15,621	2	44,9486	9	1,56393	0,261254
Fauly	0,273	2	0,9238	9	1,32912	0,312063