Téma 3: Průzkumová analýza vícerozměrných dat

Příklad 1.: Máme k dispozici datový soubor staty1979.sta z roku 1979 o 26 evropských zemích, který obsahuje údaje o procentuálním zastoupení ekonomicky činného obyvatelstva v různých odvětvích národního hospodářství:

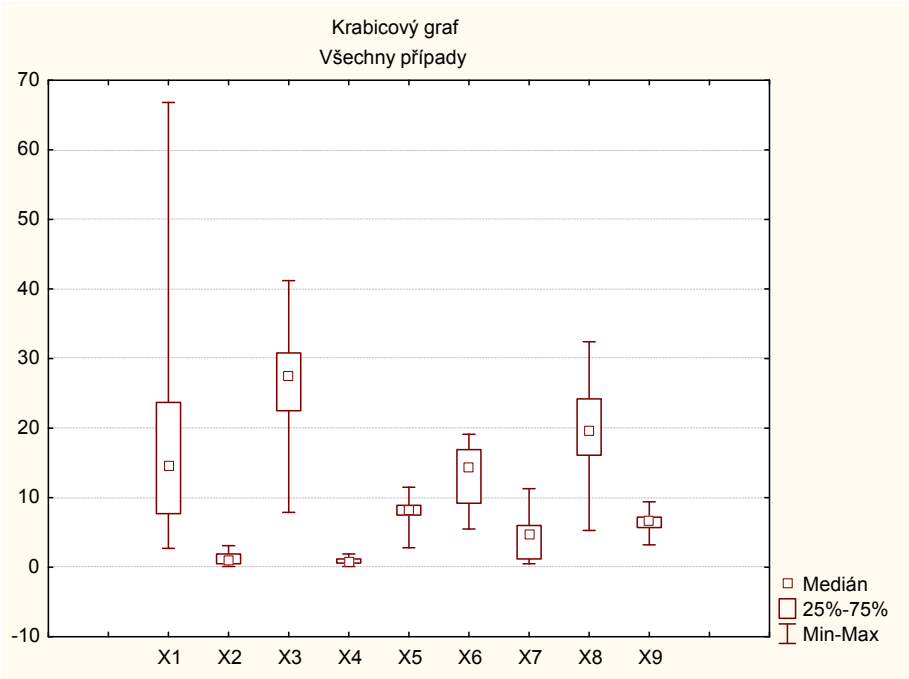
X_1 … zemědělství
X_2 … těžba
X_3 … průmyslová výroba
X_4 … energetika
X_5 … stavebnictví
X_6 … místní hospodářství
X_7 … finanční sektor
X_8 … služby
X_9 … doprava a komunikace.

	1 Stát	2 X1	3 X2	4 X3	5 X4	6 X5	7 X6	8 X7	9 X8	10 X9
1	Belgie	3,3	0,9	27,6	0,9	8,2	19,1	6,2	26,6	7,2
2	Dánsko	9,2	0,1	21,8	0,6	8,3	14,2	6,5	32,2	7,1
3	Francie	10,8	0,8	27,5	0,9	8,9	16,8	6	22,6	5,7
4	Záp. Německo	6,7	1,3	35,8	0,9	7,3	14,4	5	22,5	6,1
5	Irsko	23,2	1	20,7	1,3	7,5	16,8	2,8	20,6	6,1
6	Itálie	15,9	0,6	27,6	0,5	10	18,1	1,5	20,1	5,7
7	Lucembursko	7,7	3,1	30,8	0,8	9,2	18,5	4,5	19,2	6,2
8	Nizozemsko	6,3	0,1	22,5	1	9,9	18	6,9	28,5	6,8
9	Velká Británie	2,7	1,4	30,2	1,4	6,9	16,9	5,8	28,3	6,4
10	Rakousko	12,7	1,1	31,4	1,4	8	16,8	4,9	16,7	7
11	Finsko	13	0,4	25,9	1,3	7,4	14,7	5,5	24,2	7,6
12	Řecko	41,4	0,6	17,6	0,6	8,1	11,5	2,4	11,1	6,7
13	Norsko	9	0,5	22,4	0,8	8,6	16,9	4,7	27,7	9,4
14	Portugalsko	27,8	0,3	24,5	0,6	8,4	13,3	2,7	16,7	5,7
15	Španělsko	22,9	0,8	28,5	0,7	11,5	9,7	8,5	11,9	5,5
16	Švédsko	6,1	0,4	25,9	0,8	7,2	14,4	6	32,4	6,8
17	Švýcarsko	7,7	0,2	37,8	0,8	9,5	17,5	5,3	15,5	5,7
18	Turecko	66,8	0,7	7,9	0,1	2,8	5,5	1,1	11,9	3,2
19	Bulharsko	23,6	1,9	32,3	0,6	7,9	8	0,7	18,2	6,8
20	Československo	16,5	2,9	35,5	1,2	8,7	9,2	0,9	17,9	7,2
21	Vých. Německo	4,2	2,9	41,2	1,3	7,6	11,2	1,2	22,1	8,3
22	Maďarsko	21,7	3,1	29,6	1,9	8,2	9,4	0,9	17,2	8
23	Polsko	31,1	2,5	25,7	0,9	8,4	7,5	0,9	16,1	6,9
24	Rumunsko	34,7	2,1	30,1	0,6	8,7	5,9	1,3	11,6	5
25	Sovětský svaz	23,7	1,4	25,8	0,6	9,2	6,1	0,5	23,4	9,3
26	Jugoslávie	48,7	1,5	16,8	1,1	4,9	6,4	11,3	5,3	4

Analyzujte tato data metodou hlavních komponent a znázorněte rozmístění států na ploše prvních dvou hlavních komponent.

Řešení v systému STATISTICA:

Jednotlivé případy nejprve pojmenujeme názvy zemí.
Data – Správce jmen případů – Přenést jména případů z proměnné Stat, OK, OK.
Data nyní znázorníme pomocí krabicových diagramů:
Grafy – 2D Grafy – Krabicové grafy – Vícenásobný – Proměnné X1 až X9, OK, OK.



Proměnné vykazují značně rozdílnou variabilitu. Analýzu tedy založíme na výběrové korelační matici **R**:
Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza – Proměnné X1 až X9, OK – OK – Popisné statistiky – Korelační matice.

	Korelace (staty1979.sta)								
Proměnná	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	1,00	0,04	-0,67	-0,40	-0,53	-0,73	-0,22	-0,75	-0,56
X2	0,04	1,00	0,44	0,41	-0,02	-0,40	-0,44	-0,28	0,16
X3	-0,67	0,44	1,00	0,39	0,48	0,21	-0,15	0,15	0,36
X4	-0,40	0,41	0,39	1,00	0,03	0,20	0,11	0,13	0,37
X5	-0,53	-0,02	0,48	0,03	1,00	0,33	0,01	0,17	0,38
X6	-0,73	-0,40	0,21	0,20	0,33	1,00	0,36	0,57	0,17
X7	-0,22	-0,44	-0,15	0,11	0,01	0,36	1,00	0,11	-0,25
X8	-0,75	-0,28	0,15	0,13	0,17	0,57	0,11	1,00	0,56
X9	-0,56	0,16	0,36	0,37	0,38	0,17	-0,25	0,56	1,00

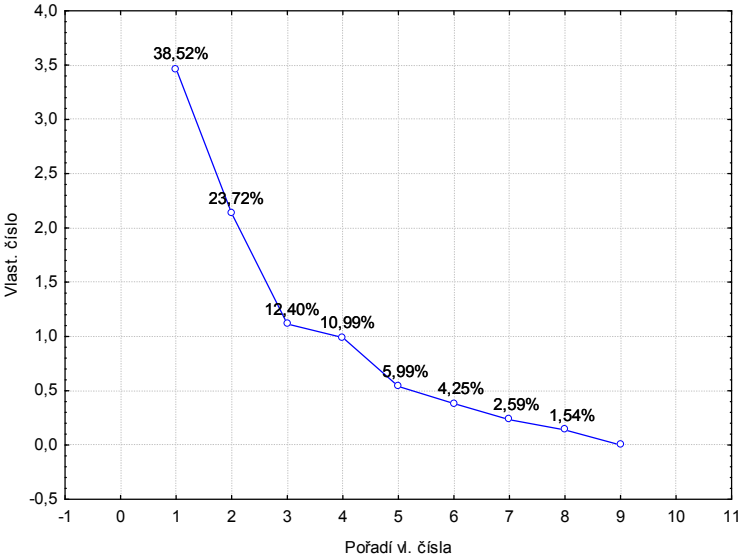
Vidíme, že některé korelační koeficienty jsou v absolutní hodnotě dostatečně velké a zřejmě tedy bude mít smysl provést analýzu hlavních komponent.

Nyní získáme vlastní čísla výběrové korelační matice a procento vysvětleného rozptylu: na záložce Základní výsledky vybereme Vlastní čísla.

Pořadí vl.č.	vl. číslo	% celk. rozptylu	Kumulativ. vl. číslo	Kumulativ. %
1	3,466490	38,51655	3,466490	38,5166
2	2,135004	23,72227	5,601494	62,2388
3	1,115581	12,39534	6,717075	74,6342
4	0,989394	10,99326	7,706468	85,6274
5	0,539211	5,99123	8,245679	91,6187
6	0,382111	4,24568	8,627790	95,8643
7	0,233226	2,59140	8,861015	98,4557
8	0,138985	1,54428	9,000000	100,0000

První hlavní komponenta tedy vysvětluje 38,52% variability obsažené v devíti sledovaných proměnných, druhá 23,72%, třetí 12,40% atd. Celkové procento variability vysvětlené prvními třemi hlavními komponentami je 74,63%.

Sestrojíme sutinový graf (scree plot): na záložce Základní výsledky vybereme Sutinový graf.



Počet hlavních komponent zvolíme tři na základě sutinového grafu, na základě vysvětleného rozptylu a na základě Kaiserova kritéria (první tři vlastní čísla jsou větší než 1). V nabídce Výsledky hlavních komponent snížíme počet faktorů na 3.

Vypočteme korelační koeficienty prvních tří hlavních komponent a původních devíti proměnných: na záložce Proměnné vybereme Korelace faktorů & proměnných.

	Korelace faktorů a proměnných (faktor. zátěže) podle korelací (staty1979.sta)		
Proměnná	Faktor 1	Faktor 2	Faktor 3
X1	0,978776	0,081725	-0,049455
X2	-0,000898	0,901105	0,216344
X3	-0,652174	0,513343	0,112868
X4	-0,474888	0,378598	0,649962
X5	-0,595263	0,073032	-0,304047
X6	-0,698213	-0,513734	0,119592
X7	-0,136193	-0,663299	0,589451
X8	-0,727506	-0,327637	-0,251642
X9	-0,684094	0,304809	-0,337074

Podívejme se rovněž na vektory souřadnic (v systému STATISTICA se jim říká faktorové souřadnice případů): na záložce Případy vybereme Faktorové souřadnice případů.

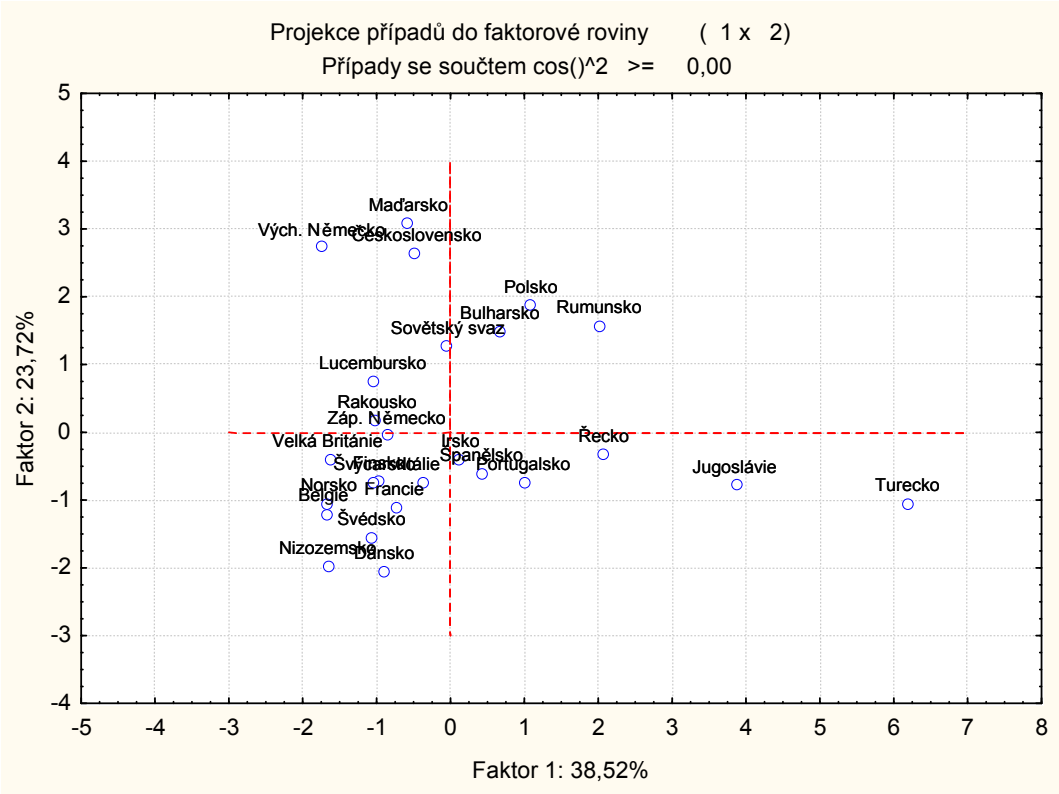
Případ	Faktorové souřadnice případů podle korelací (staty1979.sta		
	Faktor 1	Faktor 2	Faktor 3
Belgie	-1,68273	-1,20656	0,16668
Dánsko	-0,90831	-2,05598	-0,85147
Francie	-0,74050	-1,11048	0,38553
Záp. Německo	-0,85647	-0,03165	0,56466
Irsko	0,11153	-0,40400	0,53134
Itálie	-0,36366	-0,74902	-1,29050
Lucembursko	-1,04022	0,74294	0,46327
Nizozemsko	-1,65732	-1,98866	-0,08729
Velká Británie	-1,61201	-0,39776	1,35031
Rakousko	-1,01103	0,16508	1,16804
Finsko	-0,97223	-0,73166	0,54475
Řecko	2,07154	-0,33521	-0,92274
Norsko	-1,66538	-1,05092	-1,14341
Portugalsko	0,99709	-0,74259	-0,75474
Španělsko	0,43244	-0,60818	0,31825
Švédsko	-1,07387	-1,55390	-0,22815
Švýcarsko	-1,04031	-0,74707	0,28216
Turecko	6,19519	-1,04930	-0,64265
Bulharsko	0,67558	1,48159	-1,03101
Československo	-0,48005	2,63421	0,07902
Vých. Německo	-1,73669	2,73412	0,26970
Maďarsko	-0,57526	3,07981	1,09460
Polsko	1,08637	1,87264	-0,54684
Rumunsko	2,01536	1,57550	-0,48595
Sovětský svaz	-0,04779	1,26246	-2,30671
Jugoslávie	3,87872	-0,78542	3,07316

1. HK vysoce kladně koreluje s proměnnou X_1 (zemědělství) a záporně se všemi ostatními proměnnými. Tato hlavní komponenta tedy rozlišuje země na zemědělské a průmyslové. Povšimněte si, že souřadnice této hlavní komponenty jsou nejvyšší u Turecka (6,2) a Jugoslávie (3,9).

2. HK vysoce kladně koreluje s proměnnou X_2 (těžba) a podstatně slaběji s proměnnou X_3 (průmyslová výroba). Vysoké hodnoty souřadnic této hlavní komponenty najdeme u Maďarska, Východního Německa a Československa.

3. HK středně silně koreluje s proměnnou X_4 (energetika) a X_7 (finanční sektor). Nejvyšší hodnotu najdeme u Jugoslávie.

Nyní znázorníme rozmístění zemí na ploše prvních dvou hlavních komponent:
Na záložce Případy vybereme 2D graf fakt. Souřadnic příp.



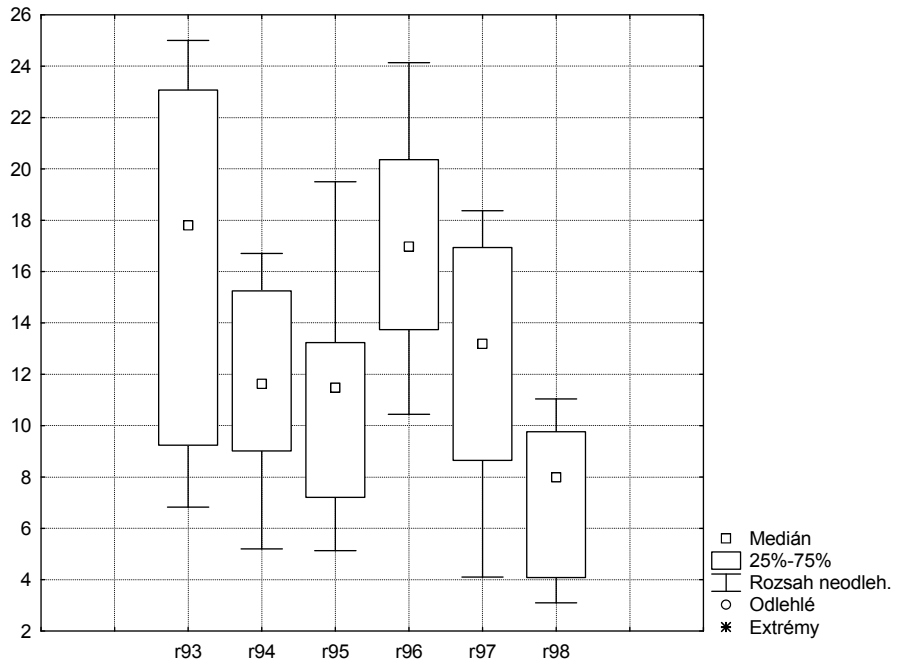
Příklad 2.: V souboru stanice.sta jsou uloženy údaje (v $\mu g/m^3$) o průměrných ročních koncentracích oxidu siřičitého v letech 1993 – 1998 na deseti brněnských měřicích stanicích: Dobrovského, Húskova, Krasová, Kroftova, Mendelova zemědělská a lesnická univerzita, Polní, Přízřenice, Skaunicové, Soběšice, Tuřany. Cílem je najít metodami shlukové analýzy skupiny stanic, které vykazují podobné rysy chování.

Datový soubor:

	1 Stanice	2 r93	3 r94	4 r95	5 r96	6 r97	7 r98
1	DOB	6,828	5,202	5,137	11,568	4,104	3,097
2	HUS	9,241	9,281	10,259	10,442	7,035	3,857
3	KRA	7,205	5,535	5,197	13,741	8,651	4,085
4	KRO	24,039	9,018	12,237	18,189	15,601	9,762
5	MZL	23,079	16,222	13,353	20,363	15,312	7,925
6	POL	25,005	14,568	10,723	15,76	11,068	4,916
7	PRI	15,874	15,251	13,241	19,435	16,943	8,081
8	SKA	14,297	9,49	7,209	14,434	10,961	8,063
9	SOB	19,728	13,772	12,943	20,948	17,564	11,039
10	TUR	22,524	16,708	19,502	24,144	18,377	11,024

Úkol 1.: Soubor stanice.sta upravte tak, aby případy 1 až 10 byly pojmenovány názvy stanic.
Návod: Data – Správce jmen případů – Přenést jména případů z proměnné Stanice, OK, OK.

Úkol 2.: Prozkoumejte proměnné r93 až r98 pomocí krabicových diagramů.
Návod: Grafy – 2D Grafy – Krabicové grafy – Vícenásobný – Proměnné r93, ..., r98, OK, OK.

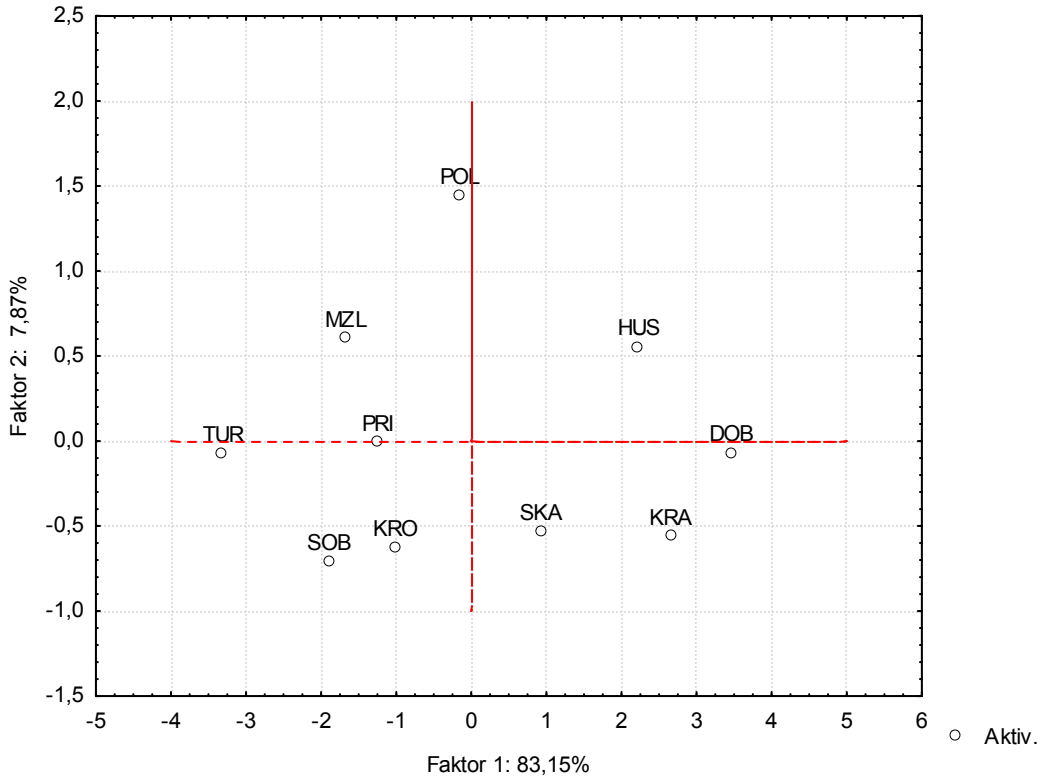


Interpretace: Z krabicových diagramů je vidět, že proměnné r93 až r98 vykazují velmi rozdílnou variabilitu. Nejvyšší variabilitu ve sledovaných deseti stanicích měly koncentrace oxidu siřičitého v roce 1993, naopak nejmenší v roce 1998.

Úkol 3.: Vzhledem k velmi rozdílné variabilitě proměnných r93 až r98 vytvořte standardizované proměnné a nadále pracujte s nimi.
Návod: Data – Standardizovat – Proměnné r93, ..., r98, OK.

	1 Stanice	2 r93	3 r94	4 r95	5 r96	6 r97	7 r98
DOB	DOB	-1,398	-1,4569	-1,3398	-1,2048	-1,7224	-1,3635
HUS	HUS	-1,0591	-0,514	-0,1653	-1,4591	-1,1255	-1,11
KRA	KRA	-1,3451	-1,3799	-1,326	-0,714	-0,7964	-1,0339
KRO	KRO	1,01924	-0,5748	0,28819	0,29058	0,61898	0,85957
MZL	MZL	0,88441	1,09043	0,54408	0,78159	0,56013	0,24685
POL	POL	1,15491	0,7081	-0,0589	-0,258	-0,3042	-0,7568
PRI	PRI	-0,1275	0,86598	0,5184	0,57199	0,89228	0,29889
SKA	SKA	-0,349	-0,4657	-0,8647	-0,5575	-0,326	0,29288
SOB	SOB	0,41376	0,5241	0,45007	0,91371	1,01875	1,2855
TUR	TUR	0,80646	1,20277	1,95397	1,63553	1,18432	1,2805

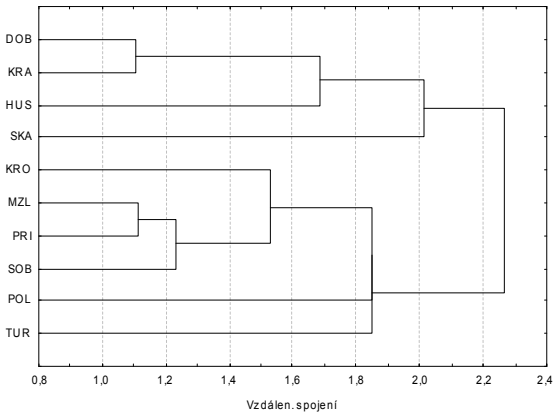
Úkol 4.: Z proměnných r93 až r98 vytvořte dvě hlavní komponenty a graficky znázorněte rozmístění stanic na ploše oprvních dvou hlavních komponent.
Návod: Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza – Proměnné r93 až r98, OK, OK – zaškrtneme 2D graf faktorových souřadnic případů.



Interpretace: Z rozmístění stanic na ploše prvních dvou hlavních komponent lze usoudit, že stanice DOB, KRA, HUS, SKA mohou tvořit jeden shluk, stanice KRO, SOB, PRI, TUR, MZL druhý shluk a stanice POL se chová poněkud atypicky.

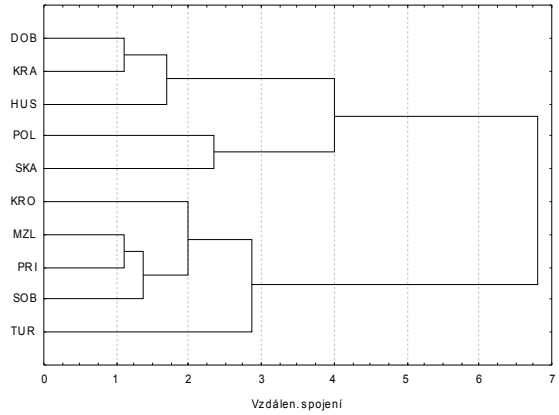
Úkol 5.: Pro standardizované proměnné r93 až r98 proveďte shlukovou analýzu s euklidovskou vzdáleností a třemi metodami: nejbližšího souseda, nejvzdálenějšího souseda a průměrné vazby. Výsledky znázorněte pomocí dendrogramu.
Návod: Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Spojování (hierarchické shlukování) – OK – Proměnné X1 – X4 – OK – na záložce Detaily vybereme Shlukovat Případy (řádky), pravidlo slučování ponecháme Jednodu-ché spojení, míru vzdálenosti ponecháme Euklidovské vzd. – OK – Horizontální graf hierarch. Stromu. Pro další dvě metody změňte Pravidlo slučování z Jednoduché spojení na Úplné spojení resp Nevážený průměr skupin dvojic.

Dendrogram pro metodu nejbližšího souseda



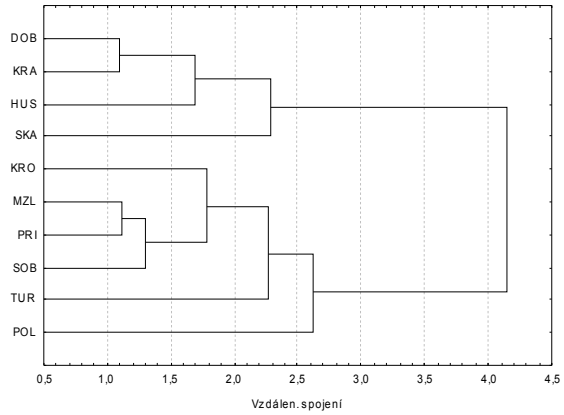
Interpretace: Stanice DOB, KRA, HUS a STA tvoří jeden shluk, stanice KRO, MZL, PRI, DOB, POL a TUR druhý shluk.

Dendrogram pro metodu nejvzdálenějšího souseda



Interpretace: Stanice DOB, KRA, HUS, POL a STA tvoří jeden shluk, stanice KRO, MZL, PRI, SOB a TUR druhý shluk.

Dendrogram pro metodu průměrné vazby



Interpretace: Stanice DOB, KRA, HUS a STA tvoří jeden shluk, stanice KRO, MZL, PRI, SOB, TUR a POL druhý shluk.

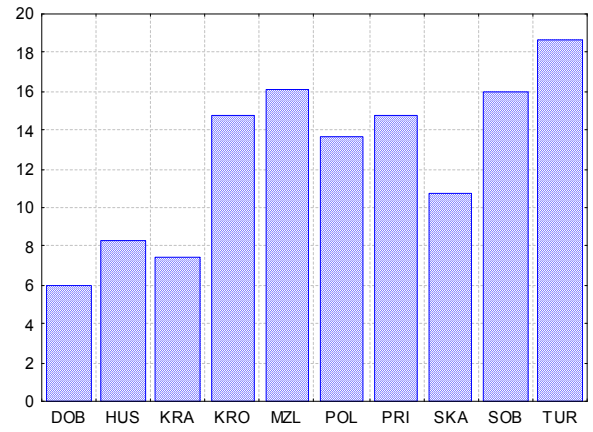
Shrneme-li výsledky všech tří metod, je zřejmé, že stanice DOB, KRA, HUS a STA zřejmě patří do jednoho shluku, zatímco stanice KRO, MZL, SOB a TUR patří do sruhého shluku. Příslušnost stanice POL k jednomu či druhému shluku není jednoznačná.

Úkol 6.: Vypočtěte a pomocí sloupkových diagramů znázorněte průměrné roční koncentrace SO_2 a směrodatné odchylky za celé sledované období pro všech deset stanic.
Návod: Je nutné se vrátit k původním nestandardizovaným hodnotám, tj. znovu načíst soubor stanice.sta a pojmenovat případy názvy stanic – viz úkol 1. Pak je zapotřebí soubor transponovat – zaměnit řádky za sloupce: Data – Transponovat – Soubor. Vymažeme 1. řádek: Upravit – Odstranit – Případy – Od případu 1 Do případu 1, OK. Pomocí Popisných statistik vypočteme průměry a směrodatné odchylky proměnných DOB až TUR.

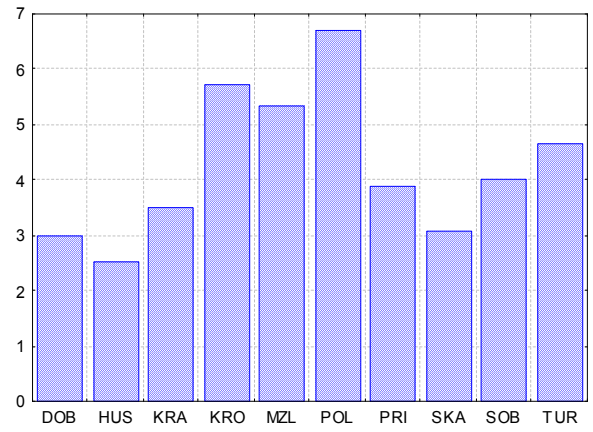
Proměnná	Popisné statistiky (stanice.sta)	
	Průměr	Sm.odch.
DOB	5,98933	3,003043
HUS	8,35250	2,513866
KRA	7,40233	3,496625
KRO	14,80767	5,707322
MZL	16,04233	5,326765
POL	13,67333	6,719292
PRI	14,80417	3,873187
SKA	10,74233	3,083617
SOB	15,99900	3,993683
TUR	18,71317	4,645334

Vytvoření sloupkových diagramů pro průměry: v Pracovním sešitě klikneme pravým tlačítkem myši na sloupek Průměr: Grafy bloku dat – Vlastní graf bloku podle sloupce –Typ grafu Sloupcové/pruhové grafy, OK. Podobně pro směrodatné odchylky.

Sloupkový diagram pro průměry



Slopupkový diagram pro sm. odchylky



Interpretace: Stanice v 1. shluku (DOB, HUS, KRA, SKA) vykazují za sledované období poměrně nízké průměrné koncentrace SO_2 (od 6 $\mu g/m^3$ po 11 $\mu g/m^3$) i malé směrodatné odchylky (od 2,5 $\mu g/m^3$ po 3,5 $\mu g/m^3$). Druhý shluk obsahuje stanice s vysokými koncentracemi (od 13 $\mu g/m^3$ po 19 $\mu g/m^3$) a velkými směrodatnými odchylkami (od 3,8 $\mu g/m^3$ po 6,8 $\mu g/m^3$).

Příklad k samostatnému řešení:
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 Tema4priklad.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

Metodami shlukové analýzy najděte skupiny hráčů podobných vlastností.

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