

## Téma č. 2.: Snížení dimenze dat metodou hlavních komponent

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

X<sub>1</sub> ... zemědělství

X<sub>2</sub> ... těžba

X<sub>3</sub> ... průmyslová výroba

X<sub>4</sub> ... energetika

X<sub>5</sub> ... stavebnictví

X<sub>6</sub> ... místní hospodářství

X<sub>7</sub> ... finanční sektor

X<sub>8</sub> ... služby

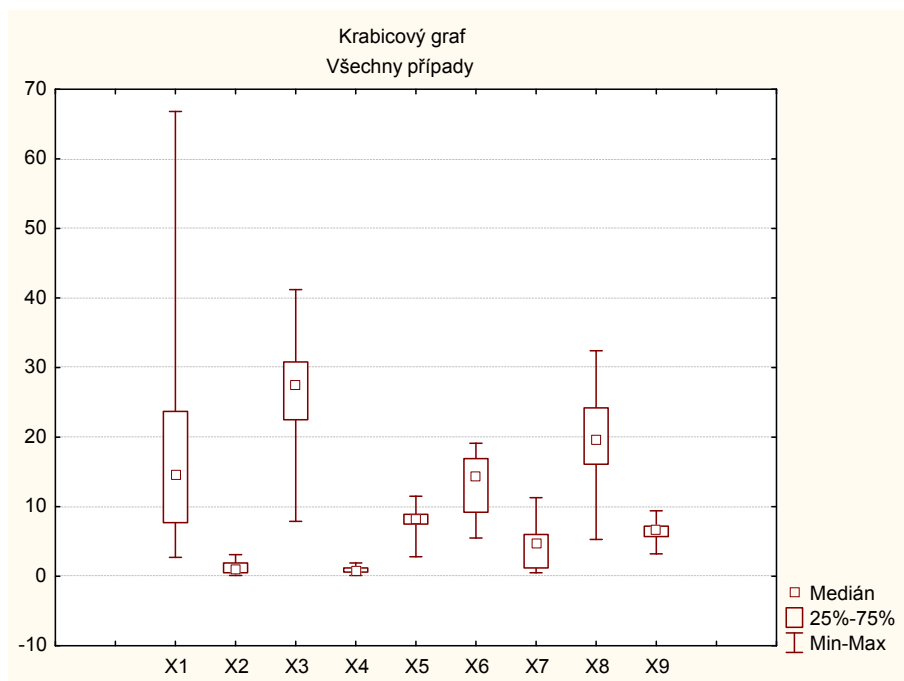
X<sub>9</sub> ... doprava a komunikace.

	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

Tento datový soubor analyzujte metodou hlavních komponent.

### Řešení v systému STATISTICA:

Data nejprve znázorníme pomocí krabicových diagramů:



Proměnné vykazují značně rozdílnou variabilitu. Analýzu tedy založíme na výběrové korelační matici **R**:

Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza  
– Proměnné X1 až X19, OK – OK – Popisné statistiky – Korelační matice.

Proměnná	Korelace (staty1979.sta)								
	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

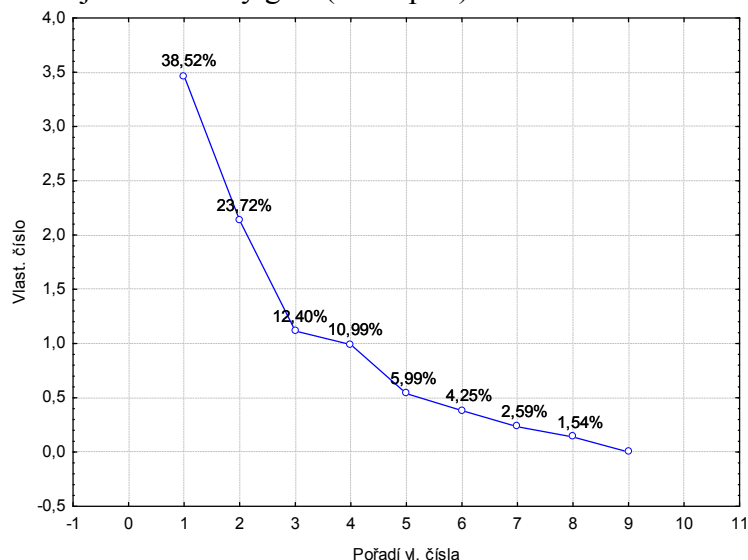
Tato korelační matice má bohužel determinant blízky 0 (říkáme, že je špatně podmíněná), nelze tedy provést Bartlettův test. Je však vidět, že některé korelační koeficienty jsou v absolutní hodnotě dostatečně velké a zřejmě tedy bude mít smysl provést analýzu hlavních komponent.

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  $m$  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.

Proměnná	Korelace faktorů a proměnných (faktor. zátěže) podle korelací (staty1979.sta)		
	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.

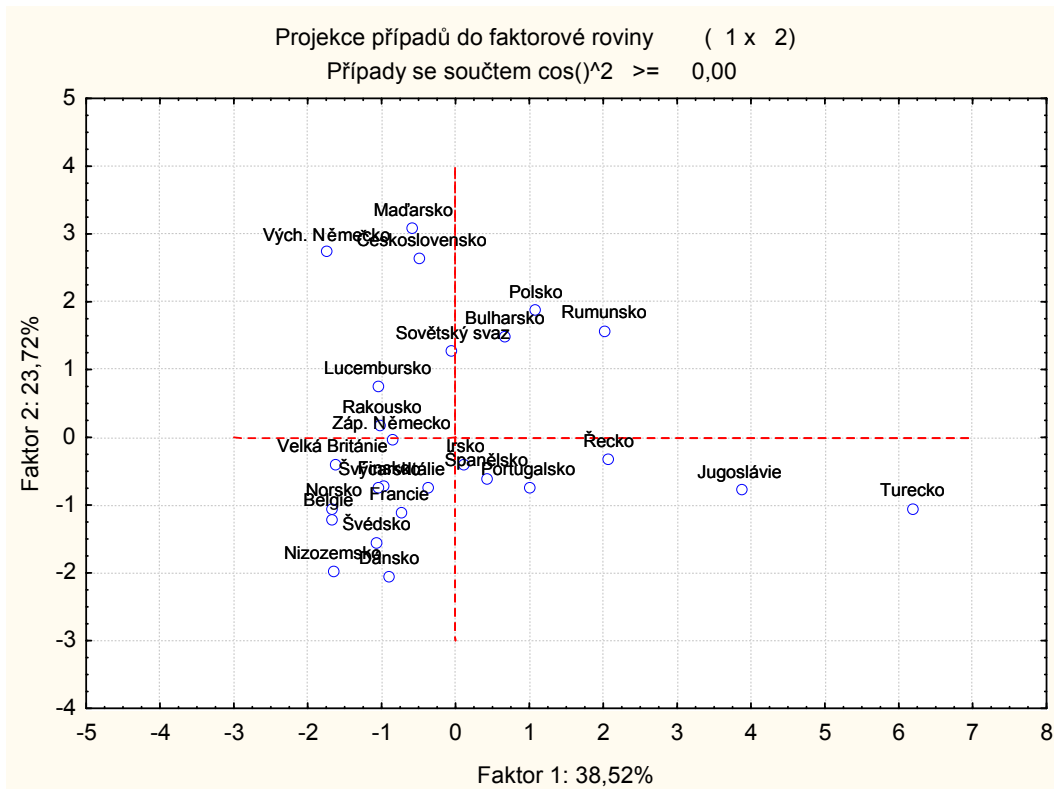
Případ	Faktorové souřadnice případů podle korelací (staty1979.sta)		
	Faktor 1	Faktor 2	Faktor 3
Belgie	-1,68273	-1,20656	0,16668
Dánsko	-0,90831	-2,05598	-0,85147
Francie	-0,74050	-1,11048	0,38553
Záp. Německo	-0,85647	-0,03165	0,56466
Irsko	0,11153	-0,40400	0,53134
Itálie	-0,36366	-0,74902	-1,29050
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).

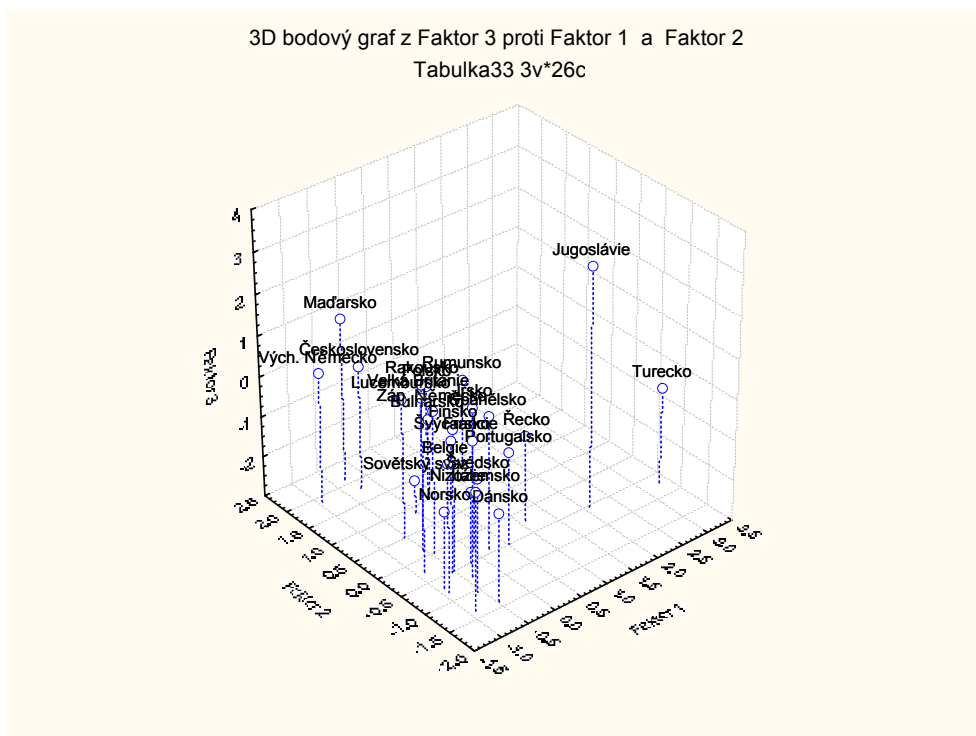
2. HK vysoce kladně koreluje s proměnnou  $X_2$  (těžba) a podstatně slaběji s proměnnou  $X_3$  (průmyslová výroba). Vysoké hodnoty souřadnic této hlavní komponenty najdeme u Maďarska, Východního Německa a Československa.

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.



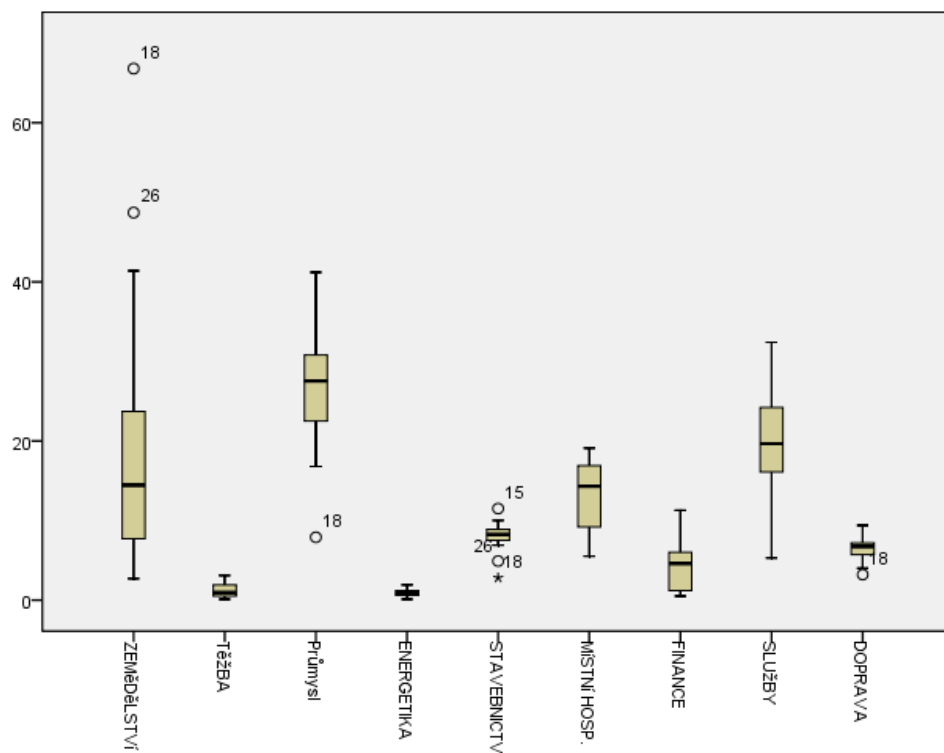
Můžeme se ještě pokusit o znázornění zemí v prostoru prvních tří hlavních komponent: přepneme se v pracovním sešitě na tabulku Faktorové souřadnice případů dle korelací. Označíme myši 3 hlavní komponenty. Klikneme pravým tlačítkem, vybereme Grafy bloku dat – Vlastní graf bloku podle sloupce – 3D XYZ grafy – Bodové grafy – Běžný – OK, 2x klikneme na pozadí grafu – Popisy bodů – zaškrtneme Zobrazovat popisy bodů.



### Řešení v systému SPSS:

Hodnoty všech 9 proměnných znázorníme pomocí krabicových diagramů.

Graphs – Legacy Dialogs – Box plot – zaškrtneme Summaries of separate variables – Define – Boxes represent x1 až x9 - OK



Výpočet výběrové korelační matice: Analyze – Correlate – Bivariate – Variables x1 až x9 – OK

Postup při provedení Bartlettova testu: Analyze – Data Reduction – Factor - Variables x1 až x9 – Descriptives – zaškrtneme KMO and Bartlett’s test of sphericity. V našem případě výsledek neobdržíme, protože korelační matice má determinant blízký 0.

Nyní získáme vlastní čísla výběrové korelační matice a procento vysvětleného rozptylu:

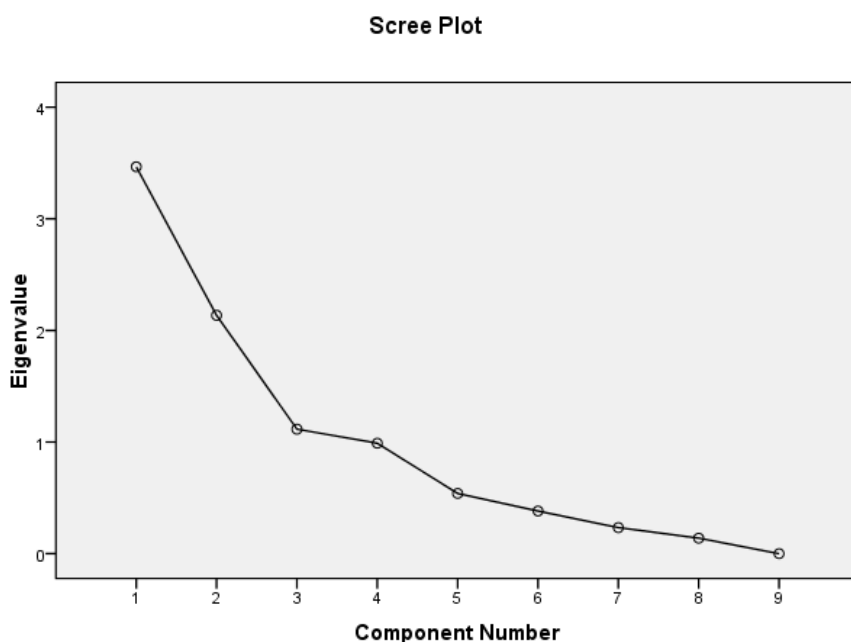
Analyze – Data Reduction – Factor - Variables x1 až x9 – OK

**Total Variance Explained**

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,466	38,517	38,517	3,466	38,517	38,517
2	2,135	23,722	62,239	2,135	23,722	62,239
3	1,116	12,395	74,634	1,116	12,395	74,634
4	,989	10,993	85,627			
5	,539	5,991	91,619			
6	,382	4,246	95,864			
7	,233	2,591	98,456			
8	,139	1,544	100,000			
9	8,077E-17	8,975E-16	100,000			

Extraction Method: Principal Component Analysis.

Chceme-li znázornit sutinový graf, ve volbě Extraction zaškrtneme Scree plot.



Ve výstupu se objeví rovněž korelační koeficienty prvních tří hlavních komponent a původních devíti proměnných:

**Component Matrix<sup>a</sup>**

	Component		
	1	2	3
ZEMĚDĚLSTVÍ	-,979	,082	-,049
TěžBA	,001	,901	,216
Průmysl	,652	,513	,113
ENERGETIKA	,475	,379	,650
STAVEBNICTVÍ	,595	,073	-,304
MÍSTNÍ HOSP.	,698	-,514	,120
FINANCE	,136	-,663	,589
SLUŽBY	,728	-,328	-,252
DOPRAVA	,684	,305	-,337

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

(Na rozdíl od systému STATISTICA dostáváme korelační koeficienty s opačnými znaménky.)

Můžeme získat rovněž reprodukovanou korelační matici a reziduální matici: ve volbě Descriptives zaškrtneme Correlation Matrix Reproduced.

Reproduced Correlations

	ZEMĚDĚLSTVÍ	TĚŽBA	PRŮMYSL	ENERGETIKA	STAVEBNICTVÍ	MÍSTNÍ HOSP.	FINANCE	SLUŽBY	DOPRAVA	
Reproduced Correlation	ZEMĚDĚLSTVÍ	,967 <sup>a</sup>	,062	-,602	-,466	-,562	-,731	-,217	-,726	-,628
	TĚŽBA	,062	,859 <sup>a</sup>	,488	,482	,001	-,436	-,470	-,349	,202
	PRŮMYSL	-,602	,488	,702 <sup>a</sup>	,577	,391	,205	-,185	,278	,565
	ENERGETIKA	-,466	,482	,577	,791 <sup>a</sup>	,113	,215	,197	,058	,221
	STAVEBNICTVÍ	-,562	,001	,391	,113	,452 <sup>a</sup>	,342	-,147	,486	,532
	MÍSTNÍ HOSP.	-,731	-,436	,205	,215	,342	,766 <sup>a</sup>	,506	,646	,281
	FINANCE	-,217	-,470	-,185	,197	-,147	,506	,806 <sup>a</sup>	,168	-,308
	SLUŽBY	-,726	-,349	,278	,058	,486	,646	,168	,700 <sup>a</sup>	,483
	DOPRAVA	-,628	,202	,565	,221	,532	,281	-,308	,483	,675 <sup>a</sup>
Residual <sup>b</sup>	ZEMĚDĚLSTVÍ		-,026	-,070	,066	,031	-,001	-,004	-,023	,065
	TĚŽBA	-,026		-,045	-,077	-,022	,040	,026	,066	-,039
	PRŮMYSL	-,070	-,045		-,185	,092	,000	,031	-,125	-,209
	ENERGETIKA	,066	-,077	-,185		-,084	-,015	-,083	,072	,154
	STAVEBNICTVÍ	,031	-,022	,092	-,084		-,011	-,153	-,314	-,147
	MÍSTNÍ HOSP.	,001	,040	,000	-,015	-,011		-,146	-,078	-,106
	FINANCE	-,004	,026	,031	-,083	,153	-,146		-,054	,057
	SLUŽBY	-,023	,066	-,125	,072	-,314	-,078	-,054		,081
	DOPRAVA	,065	-,039	-,209	,154	-,147	-,106	,057	,081	

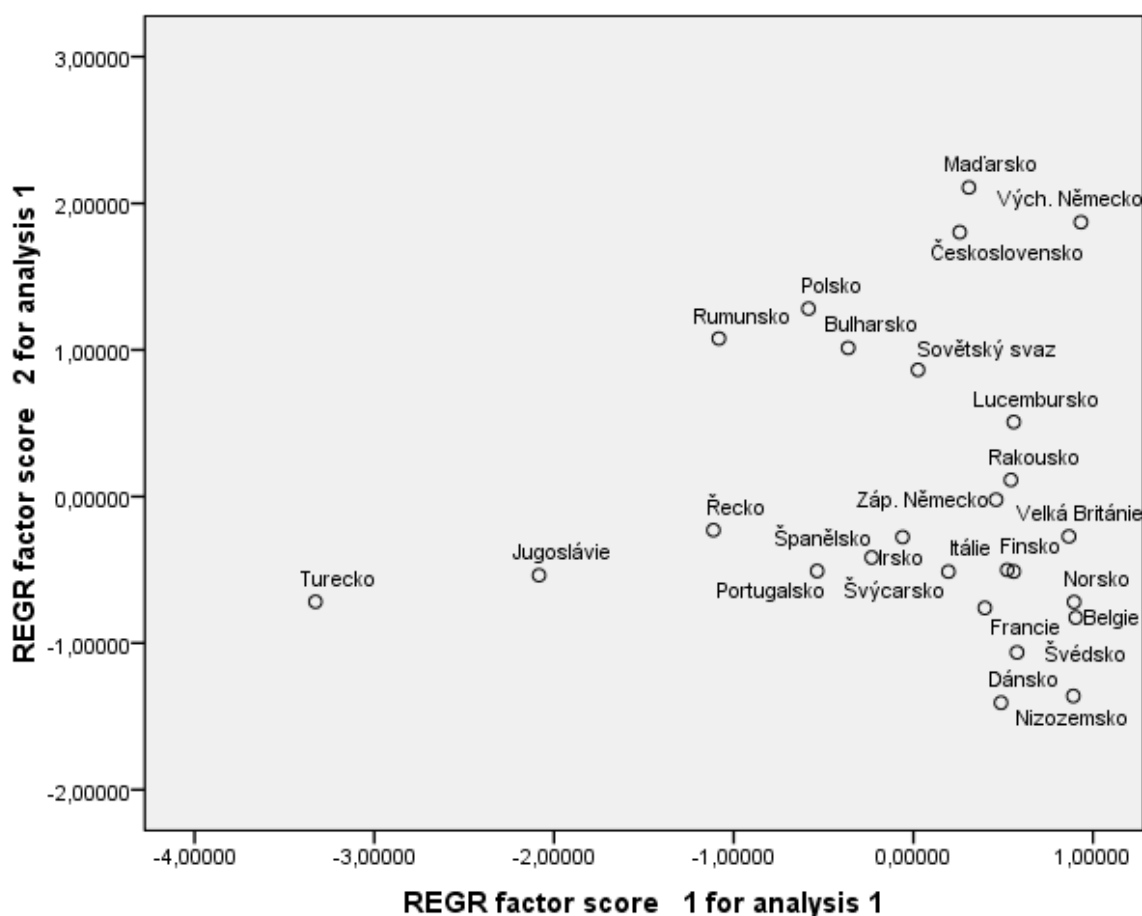
Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 22 (61,0%) nonredundant residuals with absolute values greater than 0.05.

Znázorníme rozmístění zemí na ploše prvních dvou hlavních komponent: ve volbě Scores zaškrtneme Save as variables. K datovému souboru se přidají tři nové proměnné - vektory souřadnic, nazvané FAC1\_1, FAC\_2\_1, FAC\_3\_1.

Graphs – Legacy Dialogs – Scatter/Dot – ponecháme Simple Scatter – Define – Y Axis factor score 2, X Axis factor score 1, Label Cases by stat, Options - zaškrtneme Display chart with case labels.





### **Příklad k samostatnému řešení:**

Datový soubor lide.sta obsahuje následující údaje o 32 náhodně vybraných osobách:

Sex (1 muž, 2 žena)

Věk (věk osoby v dosažených letech)

Výška (výška osoby v cm)

Hmotnost (hmotnost osoby v kg)

BMI (Body Mass Index se počítá podle vzorce  $BMI = \frac{\text{hmotnost (v kg)}}{\text{výška}^2 \text{ (v m}^2\text{)}}$ . Osoby, které mají

BMI pod 18,5, trpí podvýživou, BMI mezi 18,5 a 25 ukazuje na normální stav, hodnoty mezi 25 a 30 svědčí o nadváze a hodnoty nad 30 pak o obezitě.)

Křestním jménem osoby jsou označeny jednotlivé případy v datovém souboru.

Proveďte analýzu hlavních komponent pro tento datový soubor.