Téma č. 1.: Tabulkové a grafické zpracování vícerozměrných dat

Popis situace: V souboru staty1979 jsou uloženy sociálně ekonomické údaje o 26 evropských zemích. Data pocházejí z roku 1979, tedy z doby, kdy Evropa byla rozdělena na demokratické státy, socialistické státy a kapitalistické státy s diktaturami. Máme k dispozici údaje o procentuálním zastoupení pracovně činného obyvatelstva v různých odvětvích národního hospodářství:

X_1 … zemědělství
X_2 … těžba nerostných surovin
X_3 … průmyslová výroba
X_4 … energetika
X_5 … stavebnictví
X_6 … místní hospodářství
X_7 … finance
X_8 … služby
X_9 … doprava a komunikace

Úkol 1.: Pro všechny proměnné vytvořte tabulku číselných charakteristik (průměr, medián, minimum, maximum, směrodatná odchylka)
Návod pro systém STATISTICA: Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné 2-10, OK – Detailní výsledky – navíc zaškrtneme Medián – OK. Ve vytvořené tabulce upravíme výsledky na 1 desetinné místo.

	Popisné statistiky (staty1979.sta)					
Proměnná	N platných	Průměr	Medián	Minimum	Maximum	Sm.odch.
X1	26	19,1	14,4	2,7	66,8	15,5
X2	26	1,3	0,9	0,1	3,1	1,0
X3	26	27,1	27,6	7,9	41,2	7,0
X4	26	0,9	0,9	0,1	1,9	0,4
X5	26	8,1	8,3	2,8	11,5	1,6
X6	26	13,0	14,3	5,5	19,1	4,6
X7	26	4,0	4,6	0,5	11,3	2,8
X8	26	20,0	19,6	5,3	32,4	6,8
X9	26	6,6	6,8	3,2	9,4	1,4

Proměnné se výrazně liší jak úrovní, tak variabilitou. V průměru ve sledovaných evropských zemích pracuje nejvíce obyvatelstva v průmyslové výrobě (27,1%), nejméně v energetice (0,9%). Nejvyšší variabilitu vykazuje proměnná X_1 … procentuální podíl pracovně činného obyvatelstva v zemědělství.

Návod pro systém SPSS: Analyze – Descriptive Statistics – Descriptives – Variables X1 – X9 – OK

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
ZEMěDěLSTVÍ	26	2,70	66,80	19,1308	15,54657
TěžBA	26	,10	3,10	1,2538	,97004
Průmysl	26	7,90	41,20	27,0538	7,03353
ENERGETIKA	26	,10	1,90	,9077	,37622
STAVEBNICTVÍ	26	2,80	11,50	8,1269	1,63696
MÍSTNÍ HOSP.	26	5,50	19,10	12,9538	4,55019
FINANCE	26	,50	11,30	4,0000	2,81581
SLUŽBY	26	5,30	32,40	20,0192	6,82355
DOPRAVA	26	3,20	9,40	6,5538	1,39003
Valid N (listwise)	26				

Poznámka: Pokud bychom chtěli navíc ještě spočítat medián, museli bychom místo Descriptives zvolit Explore a dostali bychom u každé proměnné celou řadu číselných charakteristik.

Úkol 2.: Vytvořte korelační matici pro proměnné X_1 až X_9.
Návod pro systém STATISTICA: Statistika – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměnných – Proměnné 2 – 10 – OK. Na záložce Možnosti odškrtneme Včetně průměrů a sm. odch. – Výpočet.

	Korelace (staty1979.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=26 (Celé případy vynechány u ChD)								
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 nejsilnější lineární závislost (nepřímá) je mezi proměnnými X_1 (zemědělství) a X_8 (služby). Čím více pracovníků je v zemědělství, tím méně pracovníků je ve službách.

Návod pro systém SPSS: Analyze – Correlate – Bivariate – Variables X1-X9 – OK

Correlations

		ZEMěDěLSTVí	TěžBA	Průmysl	ENERGETIKA	STAVEBNICTVí	MíSTNí HOSP.	FINANCE	SLUžBY	DOPRAVA
ZEMěDěLSTVí	Pearson Correlation	1,000	,036	-,671**	-,400*	-,531**	-,731**	-,220	-,749**	-,563**
	Sig. (2-tailed)		,862	,000	,043	,005	,000	,279	,000	,003
	N	26	26	26	26	26	26	26	26	26
TěžBA	Pearson Correlation	,036	1,000	,442*	,405*	-,022	-,396*	-,444*	-,283	,164
	Sig. (2-tailed)	,862		,024	,040	,916	,045	,023	,162	,425
	N	26	26	26	26	26	26	26	26	26
Průmysl	Pearson Correlation	-,671**	,442*	1,000	,393*	,484*	,205	-,154	,153	,355
	Sig. (2-tailed)	,000	,024		,047	,012	,314	,453	,457	,075
	N	26	26	26	26	26	26	26	26	26
ENERGETIKA	Pearson Correlation	-,400*	,405*	,393*	1,000	,028	,200	,114	,130	,375
	Sig. (2-tailed)	,043	,040	,047		,891	,327	,580	,526	,059
	N	26	26	26	26	26	26	26	26	26
STAVEBNICTVí	Pearson Correlation	-,531**	-,022	,484*	,028	1,000	,330	,006	,172	,385
	Sig. (2-tailed)	,005	,916	,012	,891		,099	,975	,401	,052
	N	26	26	26	26	26	26	26	26	26
MíSTNí HOSP.	Pearson Correlation	-,731**	-,396*	,205	,200	,330	1,000	,360	,568**	,174
	Sig. (2-tailed)	,000	,045	,314	,327	,099		,071	,002	,395
	N	26	26	26	26	26	26	26	26	26
FINANCE	Pearson Correlation	-,220	-,444*	-,154	,114	,006	,360	1,000	,114	-,251
	Sig. (2-tailed)	,279	,023	,453	,580	,975	,071		,578	,216
	N	26	26	26	26	26	26	26	26	26
SLUžBY	Pearson Correlation	-,749**	-,283	,153	,130	,172	,568**	,114	1,000	,564**
	Sig. (2-tailed)	,000	,162	,457	,526	,401	,002	,578		,003
	N	26	26	26	26	26	26	26	26	26
DOPRAVA	Pearson Correlation	-,563**	,164	,355	,375	,385	,174	-,251	,564**	1,000
	Sig. (2-tailed)	,003	,425	,075	,059	,052	,395	,216	,003	
	N	26	26	26	26	26	26	26	26	26

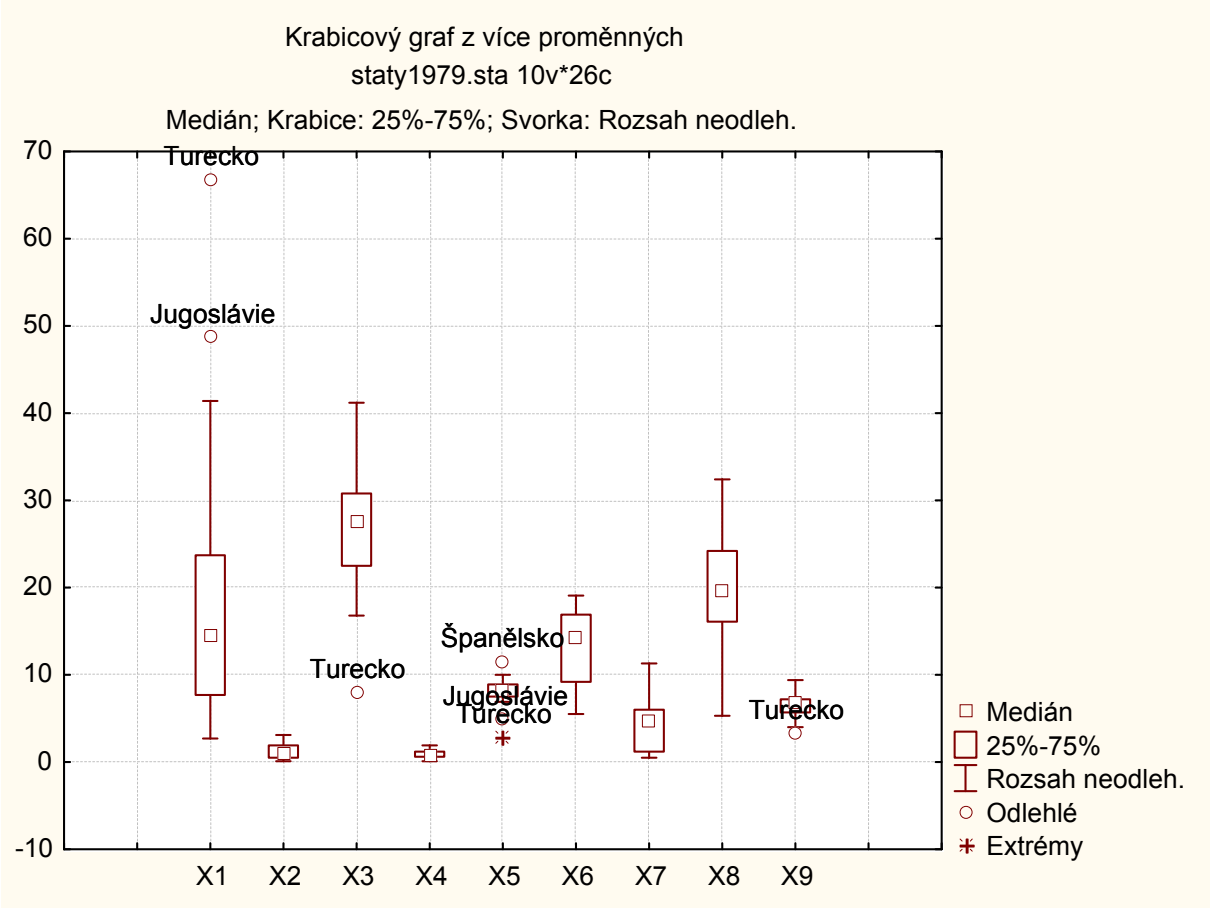
**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

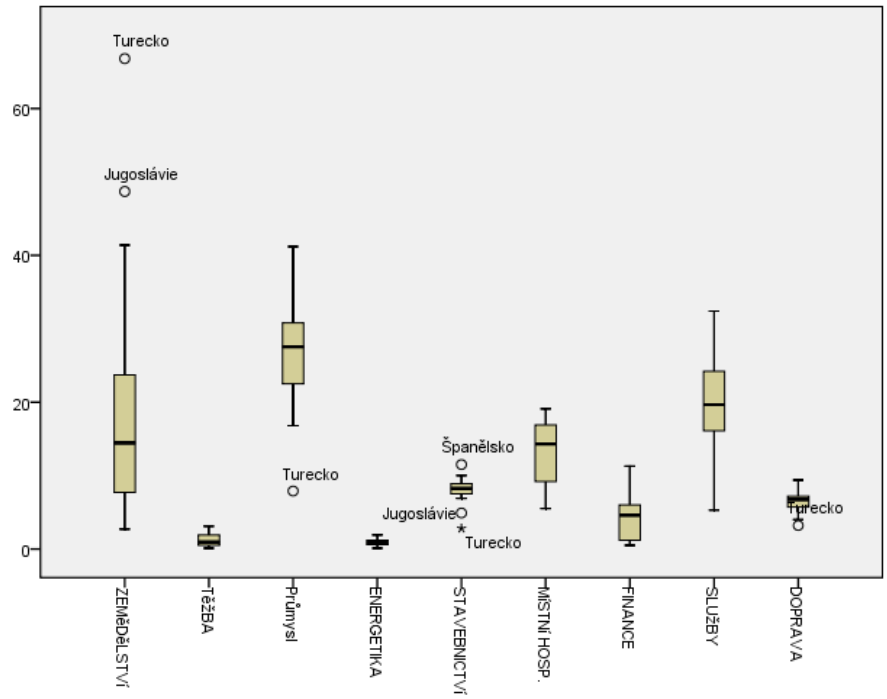
Úkol 3.: Vytvořte matici euklidovských vzdáleností pro sledovaných 26 zemí.
Návod pro systém STATISTICA: Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Spojování (hierarchické shlukování) – OK – Proměnné 2- 10 – OK – na záložce Detaily vybereme Shlukovat Případy (řádky) – OK – na záložce Detaily vybereme Matice vzdáleností.
Matice vzdáleností je příliš velká, nebudeme ji zde uvádět. Poznamenáme pouze, že největší euklidovská vzdálenost (72,2) je mezi Východním Německem a Tureckem. Naopak nejmenší euklidovská vzdálenost (4,2) je mezi Belgií a Velkou Británií..

Návod pro systém SPSS: Analyze – Classify – Hierarchical Cluster – Variables X1- X9 – Label Cases by stát - Method – Measure Euclidean distance – Continue – Statistics – zaškrtneme Proximity matrix – Continue – OK

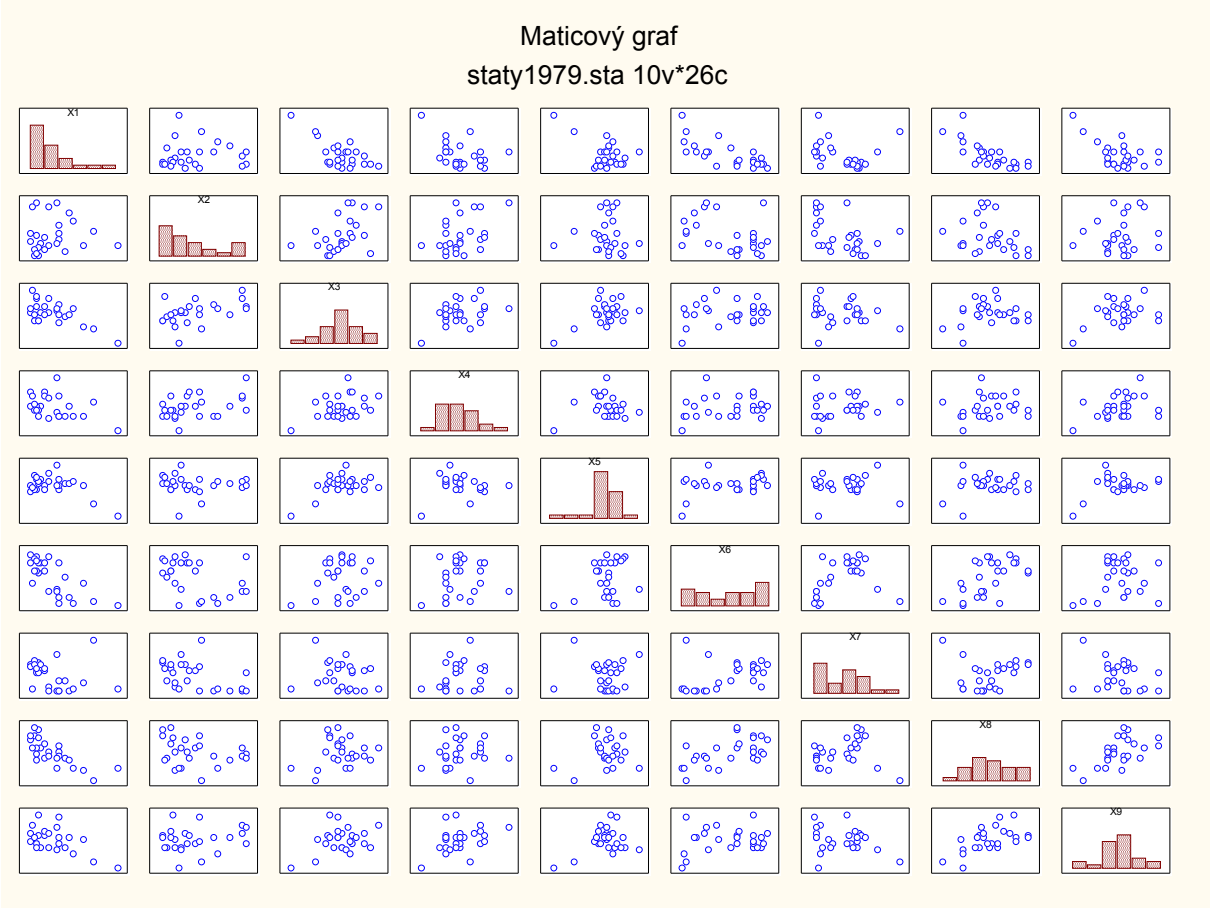
Úkol 4.: Pomocí krabicového diagramu zjistěte, zda proměnné X_1 až X_9 obsahují odlehlá či extrémní pozorování. Pokud ano, zjistěte názvy zemí, kterým tato pozorování náleží.
Návod pro systém STATISTICA: Grafy – 2D Grafy – Krabicové grafy – zvolíme Vícenásobný – Proměnné – Závislé proměnné 2 – 10 – OK. 2x klikneme na některou z odlehlých hodnot proměnné X_1, otevře se okno Rozložení grafu, vybereme záložku Popisy bodů a zaškrtneme Zobrazovat popisy bodů – OK. Podobně postupujeme u dalších proměnných.



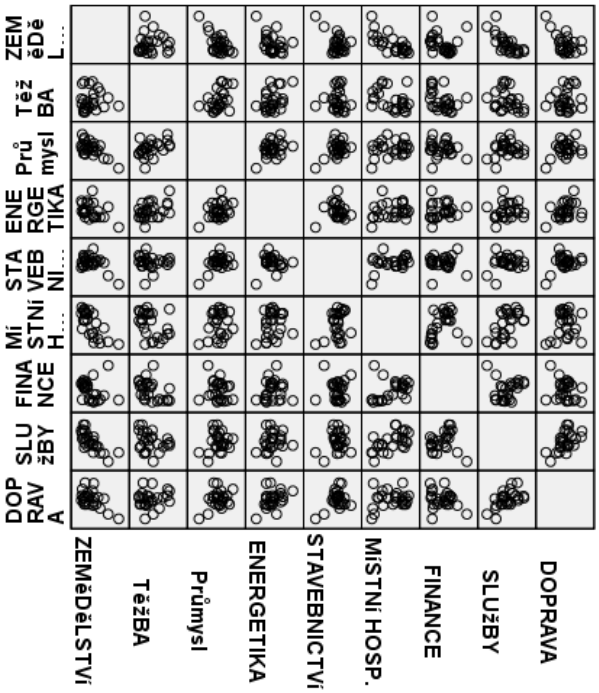
Návod pro systém SPSS: Graphs – Legacy Dialogs – Boxplot – zaškrtneme Data in Chart are Summaries of separate variables – Define – Boxes Represent X1 – X9, Label cases by stát, OK



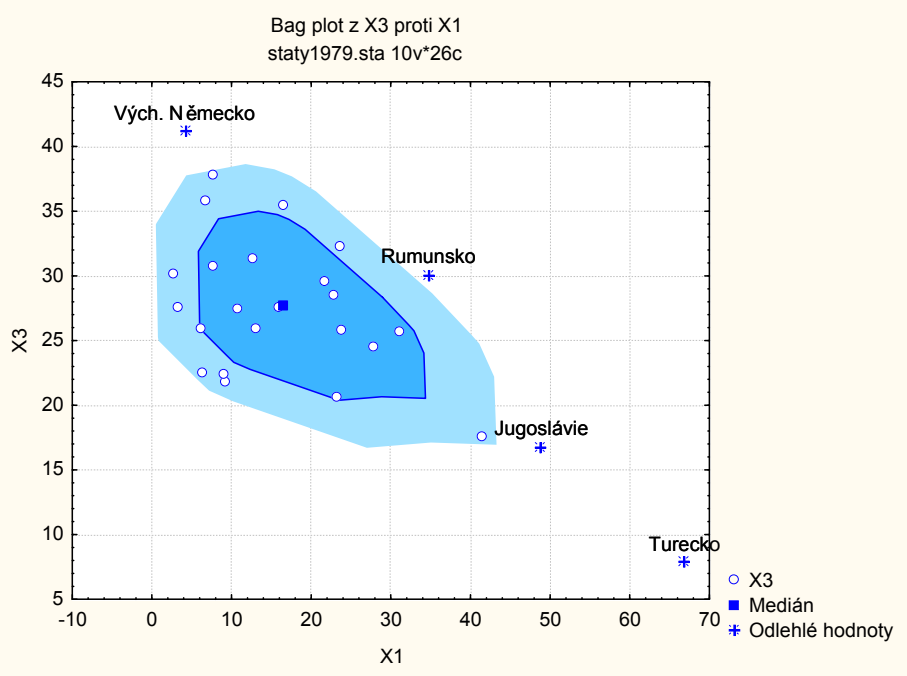
Úkol 5.: Pro proměnné X_1 až X_9 vytvořte maticový graf.
Návod pro systém STATISTICA: Grafy – Maticové grafy – Proměnné 2 – 10, OK



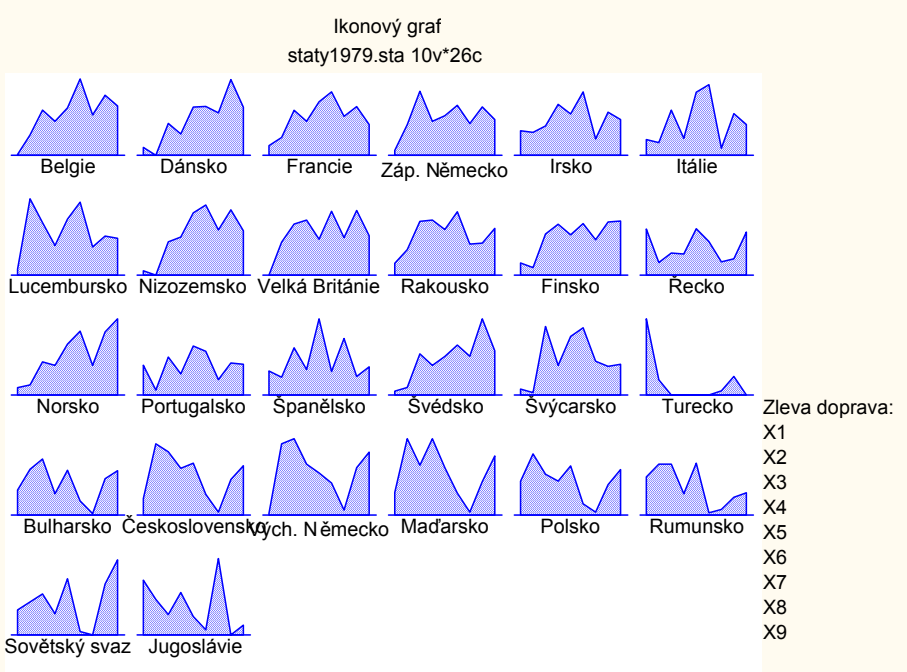
Návod pro systém SPSS: Graphs – Legacy Dialogs – Scatter/dot – Matrix Scatter – Define – Matrix Variables X1 – X9 – OK



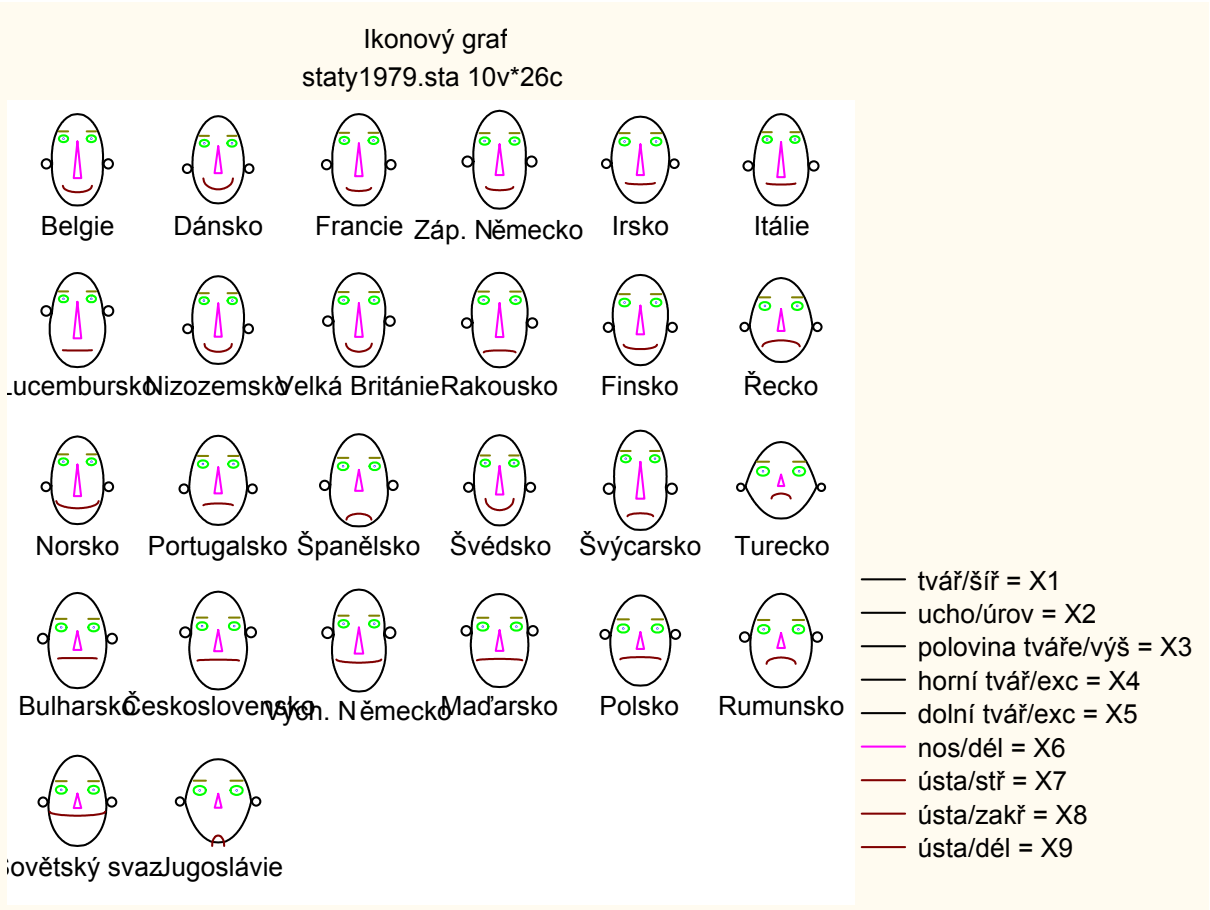
Úkol 6.: V systému STATISTICA vytvořte bag plot pro proměnné X_1 (zemědělství) a X_3 (průmysl).
Návod: Grafy – 2D Grafy – Bag Ploty – Proměnné X1a X3, OK. Ve vytvořeném grafu 2x klikneme na některou z odlehlých hodnot, otevře se okno Rozložení grafu, vybereme záložku Popisy bodů a zaškrtneme Zobrazovat popisy bodů – OK.



Úkol 7.: Pomocí systému STATISTICA vytvořte profily a Chernoffovy tváře pro proměnné X_1 až X_9.
Návod: Grafy – Ikonové grafy – Proměnné 2-10 – OK, Typ grafu Profily – Možnosti 1 – zapnout Zobrazit popisy případů, zvolit Jména případů



Pro Chernoffovy tváře zvolíme typ grafu Chernoffovy tváře.


Ikonový graf
staty1979.sta 10v*26c
Belgie
Dánsko
Francie
Záp. Německo
Irsko
Itálie
ucembursko
Nizozemsko
Velká Británie
Rakousko
Finsko
Řecko
Norsko
Portugalsko
Španělsko
Švédsko
Švýcarsko
Turecko
Bulharsko
Československo
Vých. Německo
Maďarsko
Polsko
Rumunsko
ovětský svaz
Jugoslávie
tvář/šíř = X1
ucho/úrov = X2
polovina tváře/výš = X3
horní tvář/exc = X4
dolní tvář/exc = X5
nos/dél = X6
ústa/stř = X7
ústa/zakř = X8
ústa/dél = X9