Zpracování dat
Korelace ve shlukové analýze

Jana Kuklová
Kamila Strušková

OBSAH PREZENTACE:

- Úvod – vysvětlení základních pojmů
- Regrese, regresní analýza
- Korelace, koeficienty
- Shluková analýza
- Korelace ve shlukové analýze
- Zdroje informací

Vysvětlení základních pojmů

- závislosti mezi veličinami
  - pevná: každé hodnotě veličiny odpovídá právě jedna hodnota jiné veličiny a naopak
  - volná: hodnotám jedné veličiny odpovídají různé hodnoty jiné veličiny

Vysvětlení základních pojmů

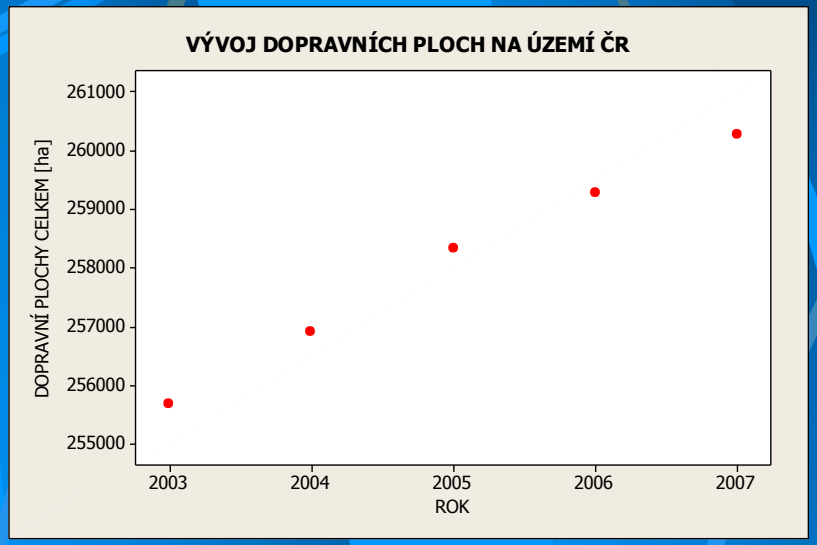
- závislosti mezi veličinami
  - jednostranná: jedna veličina (důsledek) je závislá na druhé (příčina) – regresní analýza
  - vzájemná: nelze určit, která veličina je závislá/nezávislá – korelační analýza

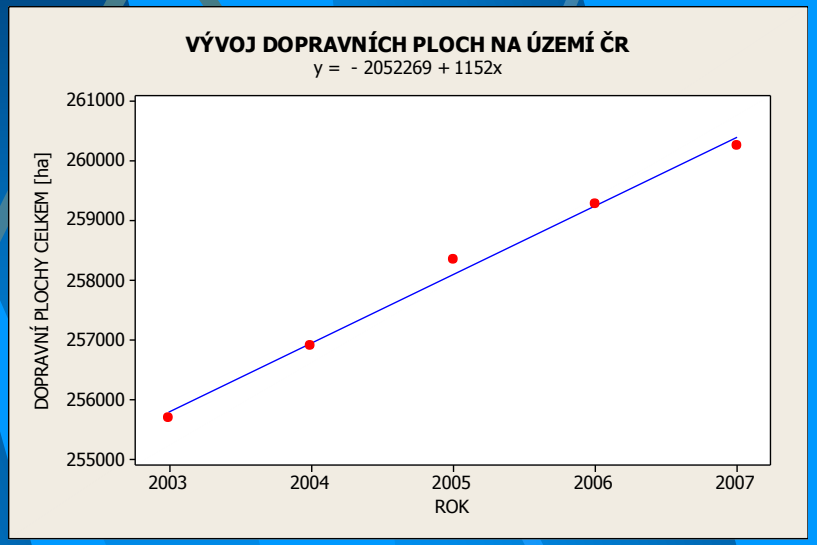
Regrese, regresní analýza

DATA
Vývoj rozlohy dopravních ploch v čase

ROK	DRÁHA	DÁLNICE	SILNICE	OSTATNÍ KOMUNIKACE	OSTATNÍ DOPRAVNÍ PLOCHY	CELKEM
	[ha]					
2003	27705	2935	69641	152775	2631	255687
2004	27772	3078	69869	153463	2719	256901
2005	27772	3212	70087	154292	2976	258339
2006	27733	3267	70330	154949	2991	259270
2007	27680	3346	70489	155731	3018	260264

Regrese, regresní analýza





lineární regrese
regresní koeficient: 99,3%

Korelační koeficient

- Udává míru linearity
- Nabívá hodnot <-1;1>
- Druhy: 1) Pearsonův korelační koeficient
  - obě proměnné jsou náhodné veličiny s normálním rozdělením
  - kardinální veličiny

  2) Spearmanův korelační koeficient
  - pro jiné než normální rozdělení
  - ordinální veličiny

  3) a další...

Korelační koeficienty

Pearsonův koeficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Spearmanův koeficient

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(i_x - i_y)^2}{n \cdot (n^2 - 1)}$$

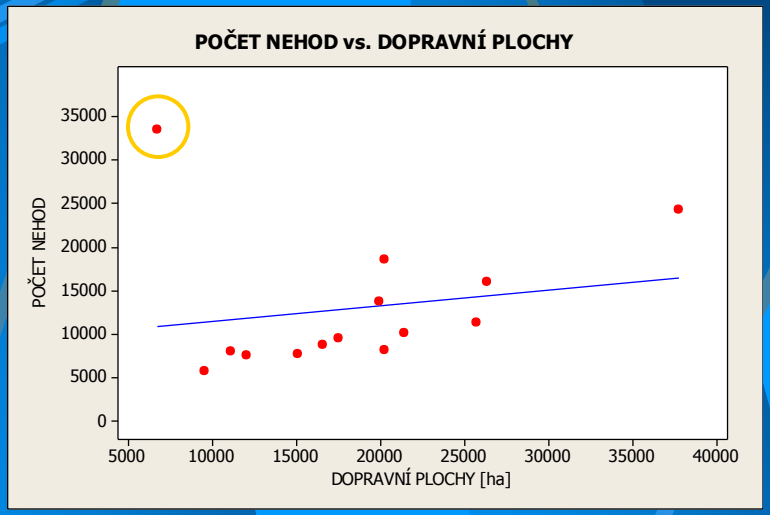
Korelace, korelační analýza

DATA

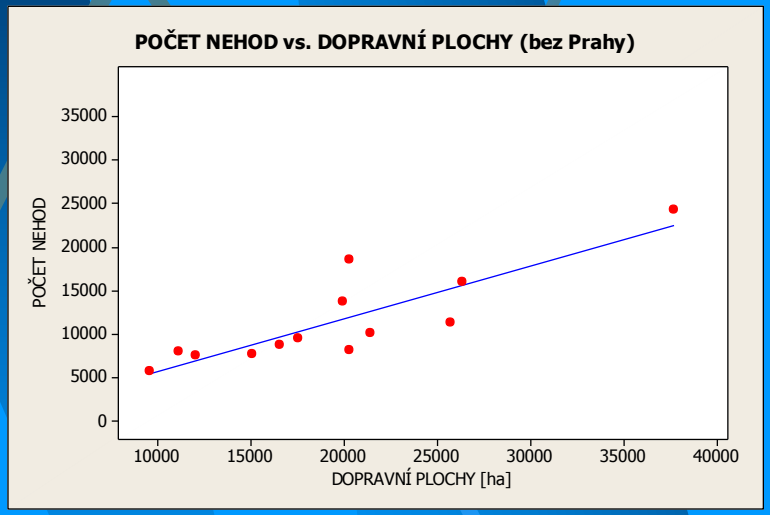
Statistické údaje jednotlivých krajů z r. 2007

Kraj	Počet obyvatel	Dopravní plochy celkem [ha]	Počet nehod
Hlavní město Praha	1201827	37699	24254
Středočeský kraj	633264	25679	11343
Jihočeský kraj	561074	21406	10151
Plzeňský kraj	307449	9571	5680
Karlovarský kraj	831180	19949	13650
Ústecký kraj	433948	11126	7993
Liberecký kraj	552212	16539	8696
Královéhradecký kraj	511400	15096	7747
Pardubický kraj	513677	20270	8086
Vysočina	1140534	26332	16022
Jihomoravský kraj	641791	17529	9545
Olomoucký kraj	590780	12053	7481
Zlínský kraj	1249290	20255	18604
Moravskoslezský kraj	1201827	37699	24254

Korelace, korelační analýza



Pearsonův korelační koeficient: 0,184



Pearsonův korelační koeficient: 0,852

Shluková analýza

- metoda učení bez učitele
- cíl: v dané množině objektů nalézt její podmnožiny (shluky) tak, aby si členové shluku byli navzájem podobní
- typy shlukové analýzy
  - hierarchická (průnikem dvou podmnožin je jedna z nich)
  - nehierarchická

Shluková analýza - příklad

DATA

Statistické údaje jednotlivých krajů z let 2002 – 2007

Kraj	VINÍK NEHODY (poměr počtu nehod v jednotlivých krajích)					
	Řidič motorového vozidla	Řidič nemotorového vozidla	Chodec	Závada komunikace	Lesní zvěř, domácí zvíře	Ostatní
Hlavní město Praha	0,972	0,002	0,011	0,004	0,003	0,008
Středočeský kraj	0,912	0,011	0,006	0,003	0,055	0,014
Jihočeský kraj	0,889	0,016	0,007	0,002	0,073	0,012
Plzeňský kraj	0,904	0,010	0,010	0,003	0,062	0,012
Karlovarský kraj	0,915	0,011	0,010	0,004	0,048	0,012
Ústecký kraj	0,916	0,011	0,011	0,004	0,047	0,011
Liberecký kraj	0,918	0,011	0,009	0,006	0,043	0,012
Královéhradecký kraj	0,921	0,018	0,007	0,003	0,041	0,010
Pardubický kraj	0,894	0,033	0,008	0,004	0,048	0,013
Vysočina	0,896	0,016	0,007	0,004	0,058	0,021
Jihomoravský kraj	0,922	0,016	0,012	0,003	0,034	0,013
Olomoucký kraj	0,901	0,031	0,007	0,003	0,039	0,019
Zlínský kraj	0,904	0,031	0,010	0,002	0,041	0,012
Moravskoslezský kraj	0,920	0,020	0,011	0,003	0,035	0,011

Shluková analýza – příklad

- objekty = kraje
- proměnné (znaky) = viník nehody
- 6 znaků: řidič motor.vozidla….
- komponenty = shluky (výsledek)
  - vzniknou na základě podobnosti jednotlivých krajů z hlediska počtu nehod

Standardizace dat

- za účelem souměřitelnosti znaků
- postup:

1. výpočet střední hodnoty jednotlivých znaků

$$\bar{z}_j = \frac{1}{n} \cdot \sum_{i=1}^{n} z_{ij}$$

2. výpočet směrodatné odchylky jednotlivých znaků

$$s_j = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(z_{ij} - \bar{z}_j \right)^2}$$

3. výpočet standardizovaných dat

$$x_{ij} = \frac{z_{ij} - \bar{z}_j}{s_j}$$

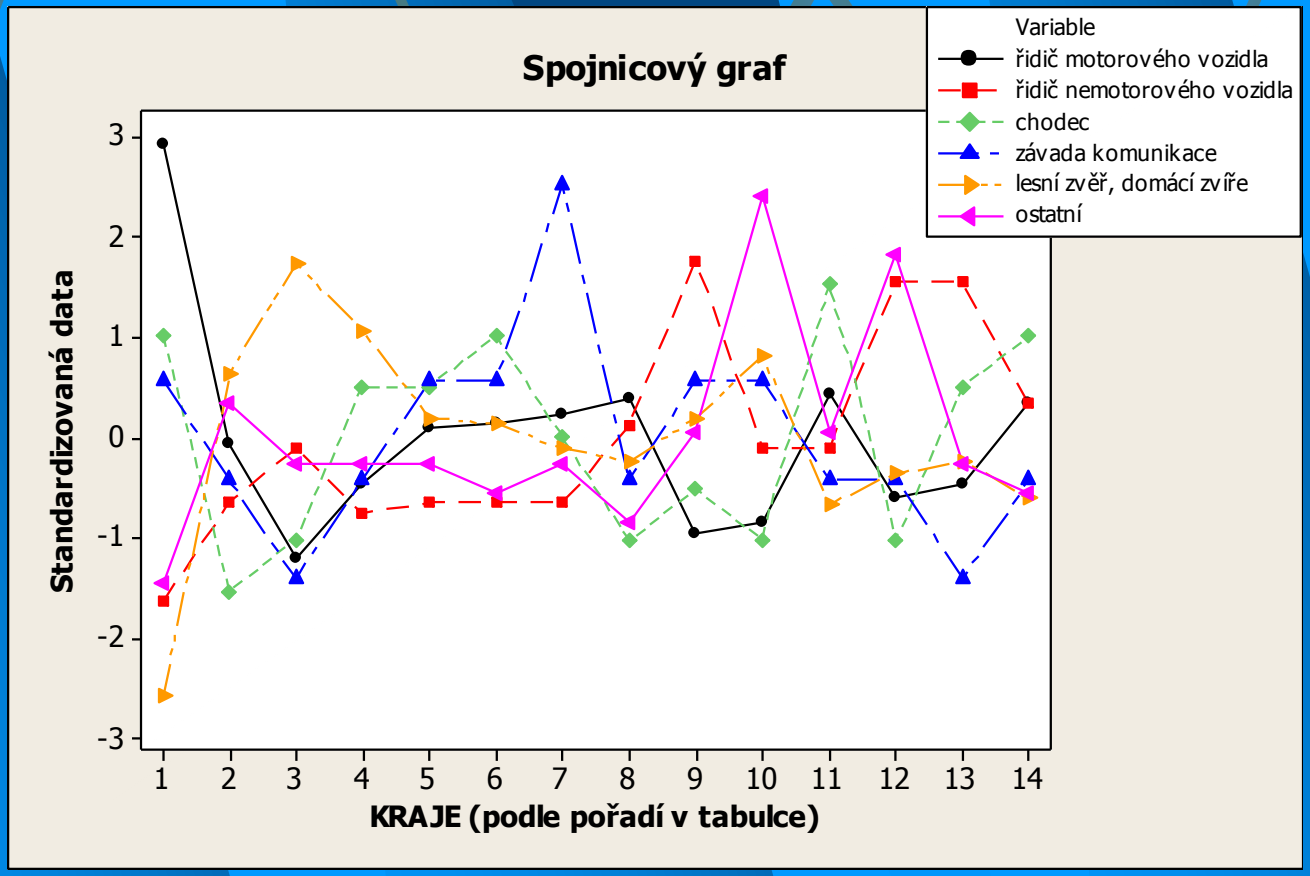
Standardizace dat

STANDARDIZOVANÁ DATA

Kraj	VINÍK NEHODY					
	Řidič motorového vozidla	Řidič nemotorového vozidla	Chodec	Závada komunikace	Lesní zvěř, domácí zvíře	Ostatní
Hlavní město Praha	2,93001	-1,63868	1,01980	0,56224	-2,57611	-1,44092
Středočeský kraj	-0,05689	-0,65077	-1,52971	-0,42168	0,62971	0,33904
Jihočeský kraj	-1,20187	-0,10193	-1,01980	-1,40559	1,73942	-0,25428
Plzeňský kraj	-0,45515	-0,76053	0,50990	-0,42168	1,06127	-0,25428
Karlovarský kraj	0,09245	-0,65077	0,50990	0,56224	0,19816	-0,25428
Ústecký kraj	0,14223	-0,65077	1,01980	0,56224	0,13651	-0,55094
Liberecký kraj	0,24180	-0,65077	-0,00000	2,53006	-0,11009	-0,25428
Královéhradecký kraj	0,39114	0,11761	-1,01980	-0,42168	-0,23339	-0,84760
Pardubický kraj	-0,95296	1,76413	-0,50990	0,56224	0,19816	0,04238
Vysočina	-0,85340	-0,10193	-1,01980	0,56224	0,81467	2,41567
Jihomoravský kraj	0,44092	-0,10193	1,52971	-0,42168	-0,66494	0,04238
Olomoucký kraj	-0,60449	1,54459	-1,01980	-0,42168	-0,35669	1,82235
Zlínský kraj	-0,45515	1,54459	0,50990	-1,40559	-0,23339	-0,25428
Moravskoslezský kraj	0,34136	0,33714	1,01980	-0,42168	-0,60329	-0,55094

Standardizace dat

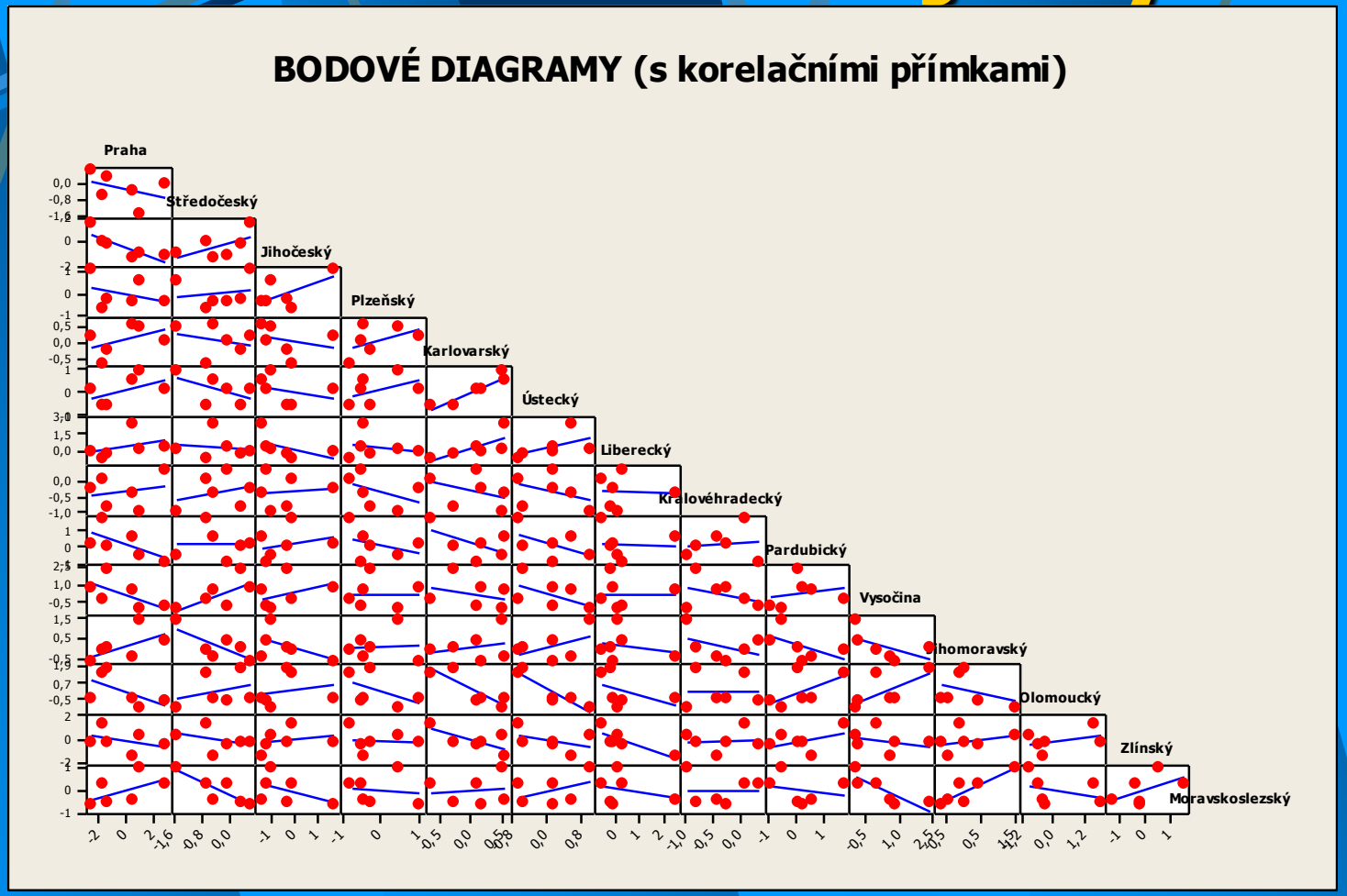
STANDARDIZOVANÁ DATA VE SPOJNICOVÉM GRAFU



Podobnost objektů

- kvantitativní vyjádření podobnosti objektů je jedním ze základních problémů shlukové analýzy
  - koeficienty asociace – pro dichotomické znaky
  - metriky (vzdálenost dvou vektorů)
    - Př.: eukleidovská, Manhattanská, Mahalanobisova…
  - korelační koeficienty – chceme-li podobnost charakterizovat přímou úměrností měřených znaků na prvním a druhém objektu
- algoritmy shlukování: alg. nejbližšího/nejvzdálenějšího souseda, centroidní alg., Wardův alg. a další

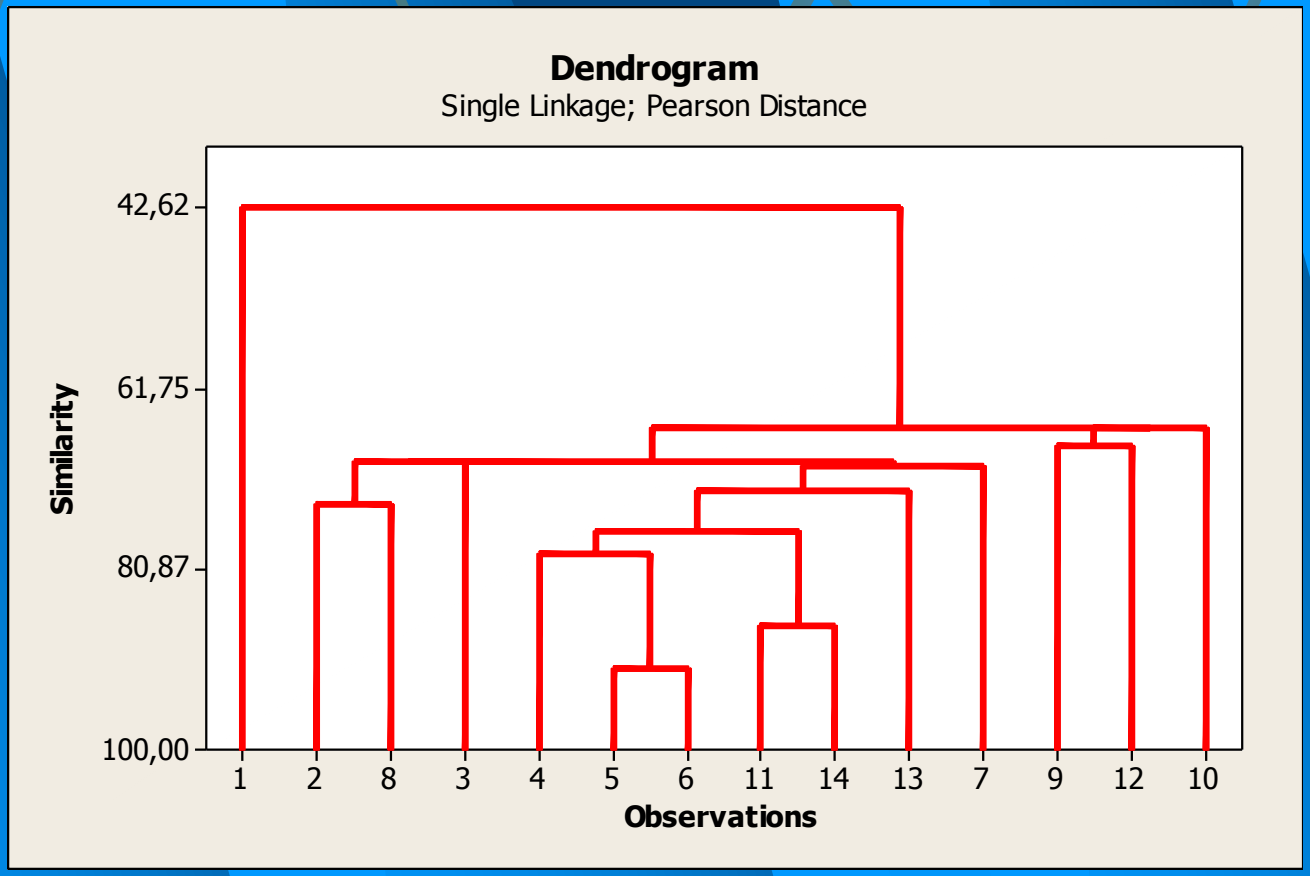
Korelace mezi objekty



Korelace mezi objekty

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Hlavní město Praha	X													
2. Středočeský kraj	-0,420	X												
3. Jihočeský kraj	-0,816	0,595	X											
4. Plzeňský kraj	-0,314	0,154	0,620	X										
5. Karlovarský kraj	0,444	-0,241	-0,288	0,482	X									
6. Ústecký kraj	0,518	-0,525	-0,353	0,466	0,935	X								
7. Liberecký kraj	0,358	-0,073	-0,494	-0,163	0,674	0,499	X							
8. Královéhradecký kraj	0,199	0,303	0,066	-0,398	-0,376	-0,368	-0,071	X						
9. Pardubický kraj	-0,646	-0,028	0,264	-0,352	-0,565	-0,532	-0,058	0,158	X					
10. Vysočina	-0,636	0,670	0,376	0,013	-0,274	-0,541	0,006	-0,339	0,234	X				
11. Jihomoravský kraj	0,572	-0,760	-0,518	0,060	0,245	0,499	-0,242	-0,388	-0,516	-0,576	X			
12. Olomoucký kraj	-0,573	0,308	0,212	-0,487	-0,867	-0,914	-0,427	0,008	0,601	0,652	-0,345	X		
13. Zlínský kraj	-0,316	-0,387	0,242	-0,091	-0,667	-0,377	-0,797	0,092	0,450	-0,285	0,309	0,401	X	
14. Moravskoslezský kraj	0,542	-0,858	-0,466	-0,092	0,070	0,404	-0,280	-0,023	-0,213	-0,840	0,872	-0,346	0,547	X

Dendrogram


Dendrogram
Single Linkage; Pearson Distance
Similarity
42,62
61,75
80,87
100,00
1
2
8
3
4
5
6
11
14
13
7
9
12
10
Observations

Zdroje informací

- Novovičová J.: Pravděpodobnost a matematická statistika
- http://www.czso.cz/ (webové stránky Českého statistického úřadu)
- Peña D.: Fundamentos de Estadística
- Ryan B.F., Joiner B.L., Cryer J.D.: Minitab Handbook
- http://gerstner.felk.cvut.cz/biolab/X33BMI/slides/KMeans.pdf (Kelbel J., Šilhán D.: Shluková analýza)
- http://meloun.upce.cz/docs/publication/152.pdf