Vícerozměrné metody - cvičení

Mgr. Jan Zdražil

Přínos cvičení

- prohloubení teoretických a praktických znalostí vícerozměrné analýzy dat
- schopnost zvolit a aplikovat adekvátní metodu vícerozměrné analýzy dat k dosažení požadovaných výsledků
- schopnost interpretovat výsledky získané prostřednictvím vícerozměrných metod

- konkrétní probíraná témata:
  - vizualizace a popis vícerozměrných dat
  - vícerozměrné statistické testy
  - výpočet podobností a vzdáleností ve vícerozměrném prostoru
  - výpočet a vizualizace asociačních matic
  - shluková analýza a její aplikace při analýze vícerozměrných dat
  - aplikace metod ordinační analýzy na vícerozměrná data

- doporučená literatura: https://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat--vicerozmerne-metody-pro-analyzu-dat

Požadavky k zápočtu

- maximálně 2 absence
- vypracování závěrečného projektu

Cvičení 1

Vizualizace vícerozměrných dat

Vícerozměrná data

PROMĚNNÉ

OBJEKTY (SUBJEKTY)

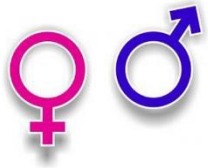
ID	Pohlaví	Věk	Váha	MMSE skóre	Objem hipokampu	…
1	muž	84	85,5	29	7030	
2	žena	25	62,0	28	6984	
3						
4						
…						

Poznámka: proměnné označovány i jako znaky, pozorování, diskriminátory, příznakové proměnné či příznaky

Anglicky označení pouze jedním termínem: feature

Typy dat - opakování

- **Kvalitativní (kategoriální) data:**
  - Binární data



  - Nominální data

  - Ordinální data

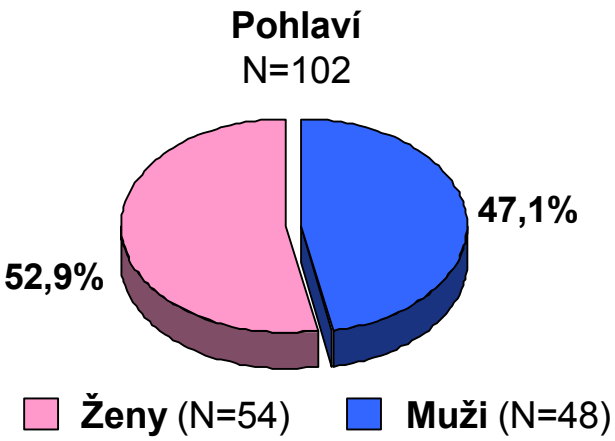
- **Kvantitativní data:**
  - Intervalová data

  - Poměrová data

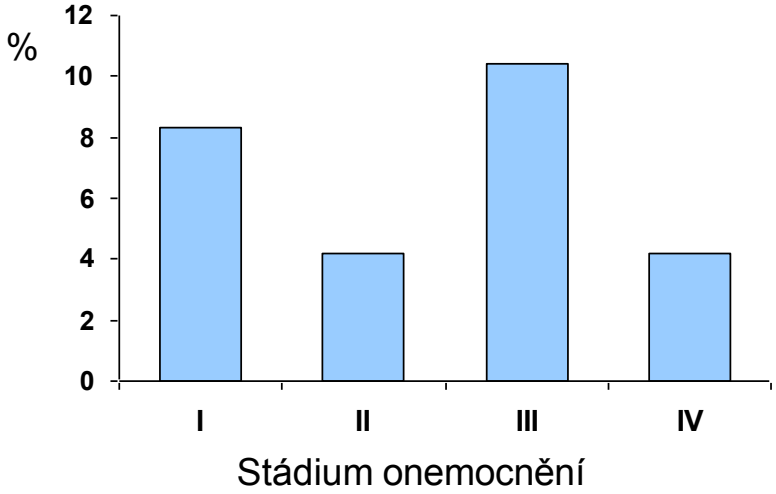
  - Spojitá data
  - Diskrétní data

Vizualizace jednorozměrných dat - opakování

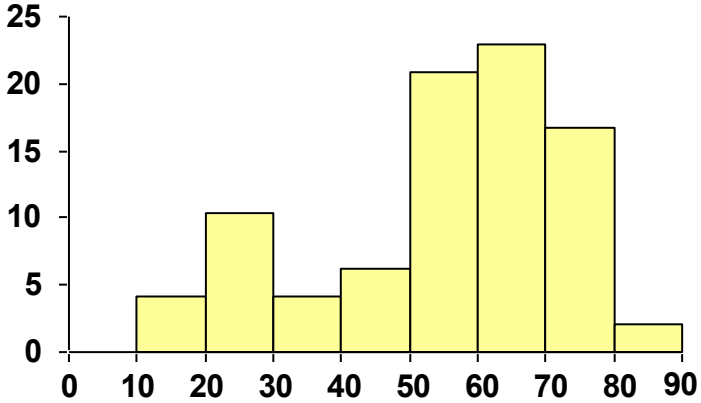
Koláčový graf



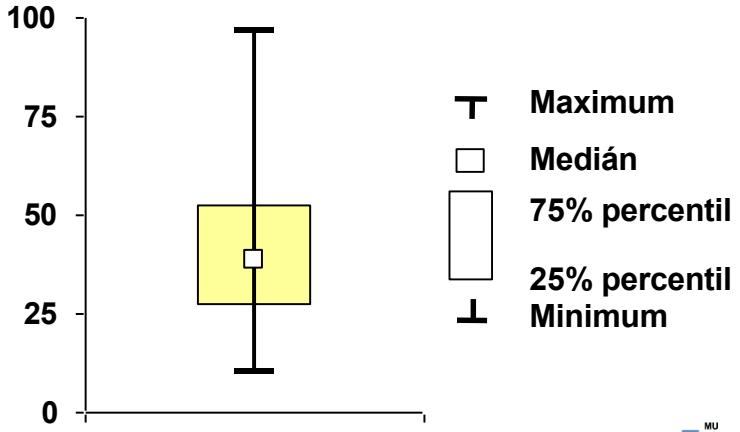
Sloupkový graf



Histogram



Krabicový graf (Box Plot)



K čemu nám může pomoci vizualizace dat?

→ ke zjištění rozložení dat, k odhalení vztahů mezi proměnnými apod.

→ odhalení problémů v datech

id	vek	pohlavi	cholesterol	vyska	vaha	obvod_pasu	obvod_boku	BMI	sys_tlak	dia_tlak
1	38	Z	4.6	164	45	60	87	16.7	120	80
2	36	Z	4.35	167	90	97	112	32.3	130	80
3	26	Z		178	70	72	94	22.1	127	80
4	25	Z	4.2	165	59	65	92	21.7	130	80
5	47	M	5.65	158		92	96	26.8	155	90
6	21	Z	6.35	172	61	69	98	20.6	135	80
7	23	Z	3.45	170	82	92	113	28.4	130	80
8	35	M	7.99	179	90	101	110	28.1	140	88
9	33	Z	4.88	167	57	70	92	20.4	140	85
10	48	Z	9.56	164	70	93	107	26.0	250	97
11	25	M	3.1	186	75	81	102	21.7	120	70
12	41	Z	10	167	62	71	101	22.2	140	90
13	29	ZZ	4.2	165	58	66	98	21.3	120	80
14	24	M	5.62	174	80	92	107	26.4	156	90
15	58	Z	7.9	164	63	73	100	23.4	135	90

K čemu nám může pomoci vizualizace dat?

→ ke zjištění rozložení dat, k odhalení vztahů mezi proměnnými apod.

→ odhalení problémů v datech

id	vek	pohlavi	cholesterol	vyska	vaha	obvod_pasu	obvod_boku	BMI	sys_tlak	dia_tlak
1	38	Z	4.6	164	45	60	87	16.7	120	80
2	36	Z	4.35	167	90	97	112	32.3	130	80
3	26	Z		178	70	72	94	22.1	127	80
4	25	Z	4.2	165	59	65	92	21.7	130	80
5	47	M	5.65	158		92	96	26.8	155	90
6	21	Z	6.35	172	61	69	98	20.6	135	80
7	23	Z	3.45	170	82	92	113	28.4	130	80
8	35	M	7.99	179	90	101	110	28.1	140	88
9	33	Z	4.88	167	57	70	92	20.4	140	85
10	48	Z	9.56	164	70	93	107	26.0	250	97
11	25	M	3.1	186	75	81	102	21.7	120	70
12	41	Z	10	167	62	71	101	22.2	140	90
13	29	ZZ	4.2	165	58	66	98	21.3	120	80
14	24	M	5.62	174	80	92	107	26.4	156	90
15	58	Z	7.9	164	63	73	100	23.4	135	90

Chybné hodnoty

Chybějící hodnoty

Odlehlé hodnoty

Problémy v datech – chybějící hodnoty

- snaha, aby v datech vůbec nenastaly
- pokud však nastanou, je silně nedoporučováno dělat každou analýzu na jinak velkém souboru (tzv. „pairwise“ odstraňování objektů) → 3 možná řešení:

1. vyloučit z analýzy všechny objekty, u nichž se vyskytla nějaká chybějící hodnota (tzv. „listwise“= „casewise” odstranění objektů):
   - pokud chybějících hodnot mnoho, zbyde pouze málo objektů
   - pozor na systematicky chybějící hodnoty – může dojít ke zkreslení výsledků analýz
   - občas vhodné odstranit proměnné s mnoha chybějícími hodnotami místo objektů, pokud proměnné nejsou důležité pro analýzu
2. definování souboru s vyplněnými „klíčovými“ proměnnými:
   - na tomto souboru provedena většina analýz
   - další analýzy dělány na podsouboru s menším počtem subjektů
3. doplnění chybějících hodnot (tzv. imputace):
   - doplnění průměrem z hodnot, které jsou pro danou proměnnou k dispozici
   - doplnění hodnot na základě regresních modelů
   - pozor! doplnění hodnot však může zkreslit výsledky analýz

Problémy v datech – odlehlé hodnoty

- k identifikaci odlehlých hodnot mohou pomoci např. tečkové, maticové či krabicové grafy
- je třeba rozlišovat:

1. **odlehlé hodnoty, které jsou způsobeny chybou** (měřících přístrojů apod.) - jsou to většinou nereálné hodnoty → je vhodné je smazat a dále s nimi zacházet jako s chybějícími hodnotami

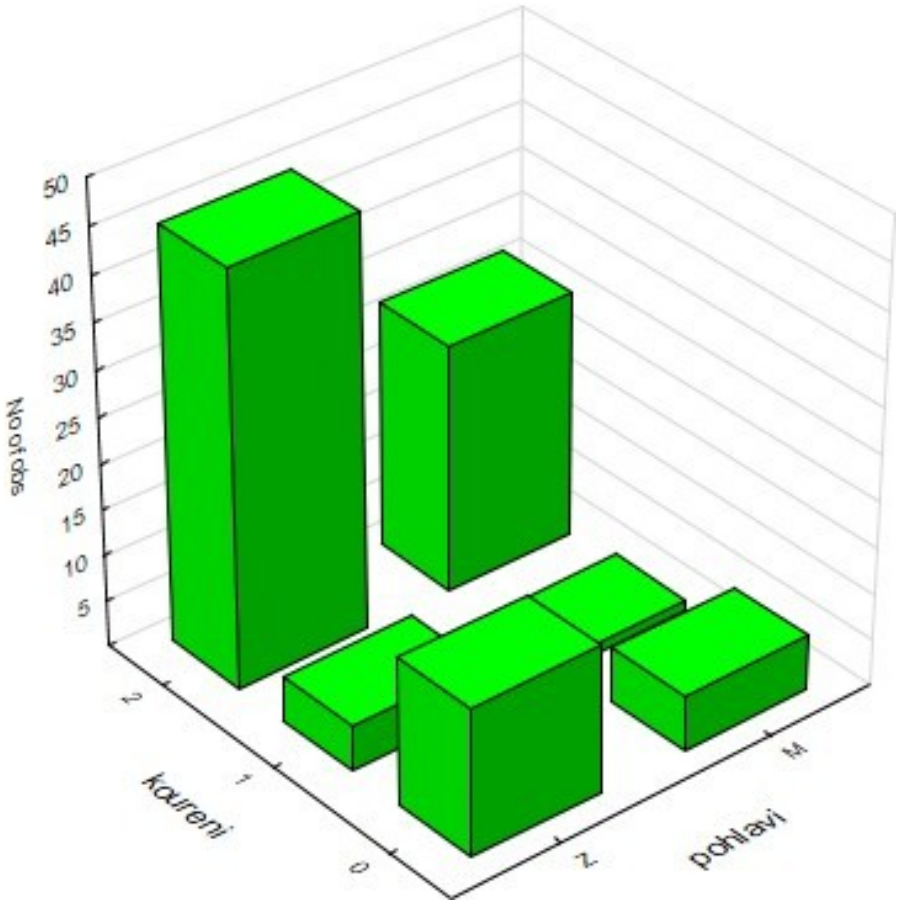
2. **odlehlé hodnoty, které jsou fyziologické** (tzn. jsou to reálné hodnoty) → je vhodné tyto hodnoty v datech ponechat, pokud je to možné a nezkreslí to analýzu a použít neparametrické metody analýzy dat
   - příklad, kdy je vhodné odlehlou hodnotu v souboru ponechat: pacienti Alzheimerovou chorobou v našem souboru mají hodnotu MMSE skóre větší než 15, jeden pacient má však hodnotu skóre 7 (je to reálná hodnota, smazáním bychom uměle snížili variabilitu)
   - příklad, kdy je nevhodné odlehlou hodnotu v souboru ponechat: chceme měřit výšku 15-letých dětí – dítě trpící nanismem měřící 80 cm by průměrnou výšku velice zkreslilo, proto ho ze souboru vyřadíme

Vizualizace vícerozměrných dat

- 3D sloupkové grafy
- dvourozměrný histogram
- maticové grafy
- krabicové grafy pro více proměnných
- ikonové (symbolové) grafy:
  - profilové sloupce
  - profily
  - paprskové (hvězdicové) grafy
  - polygony
  - pavučinové grafy
  - Chernoffovy tváře

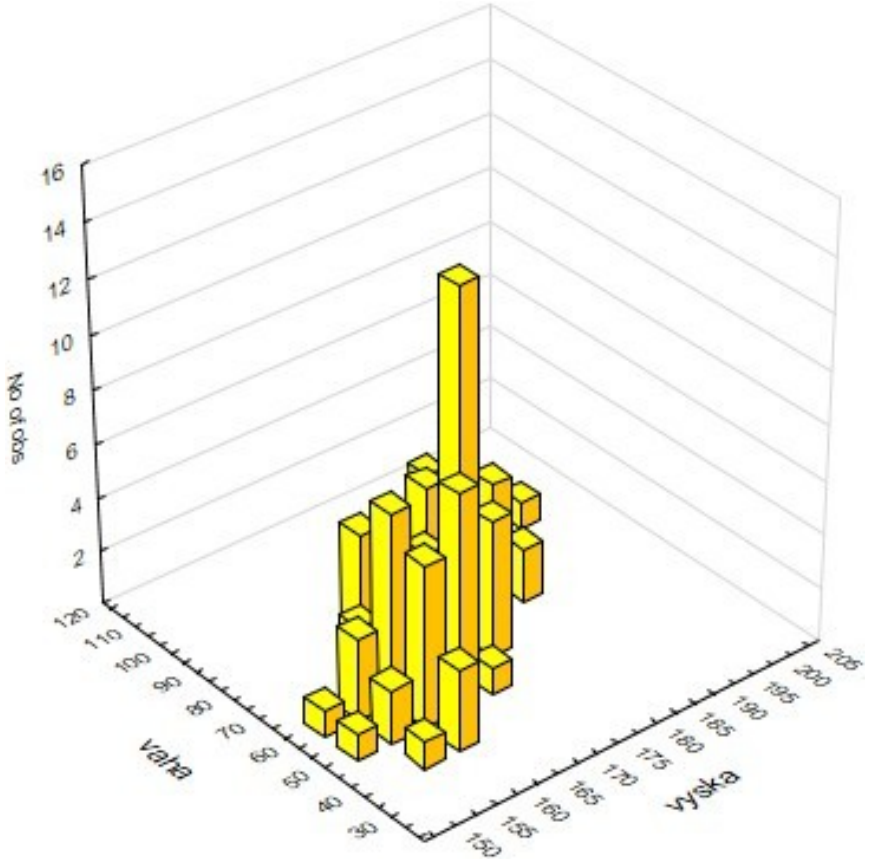
3D sloupkové grafy

- vzájemný výskyt kategorií dvou kategoriálních proměnných
- v softwaru R: Barplot()



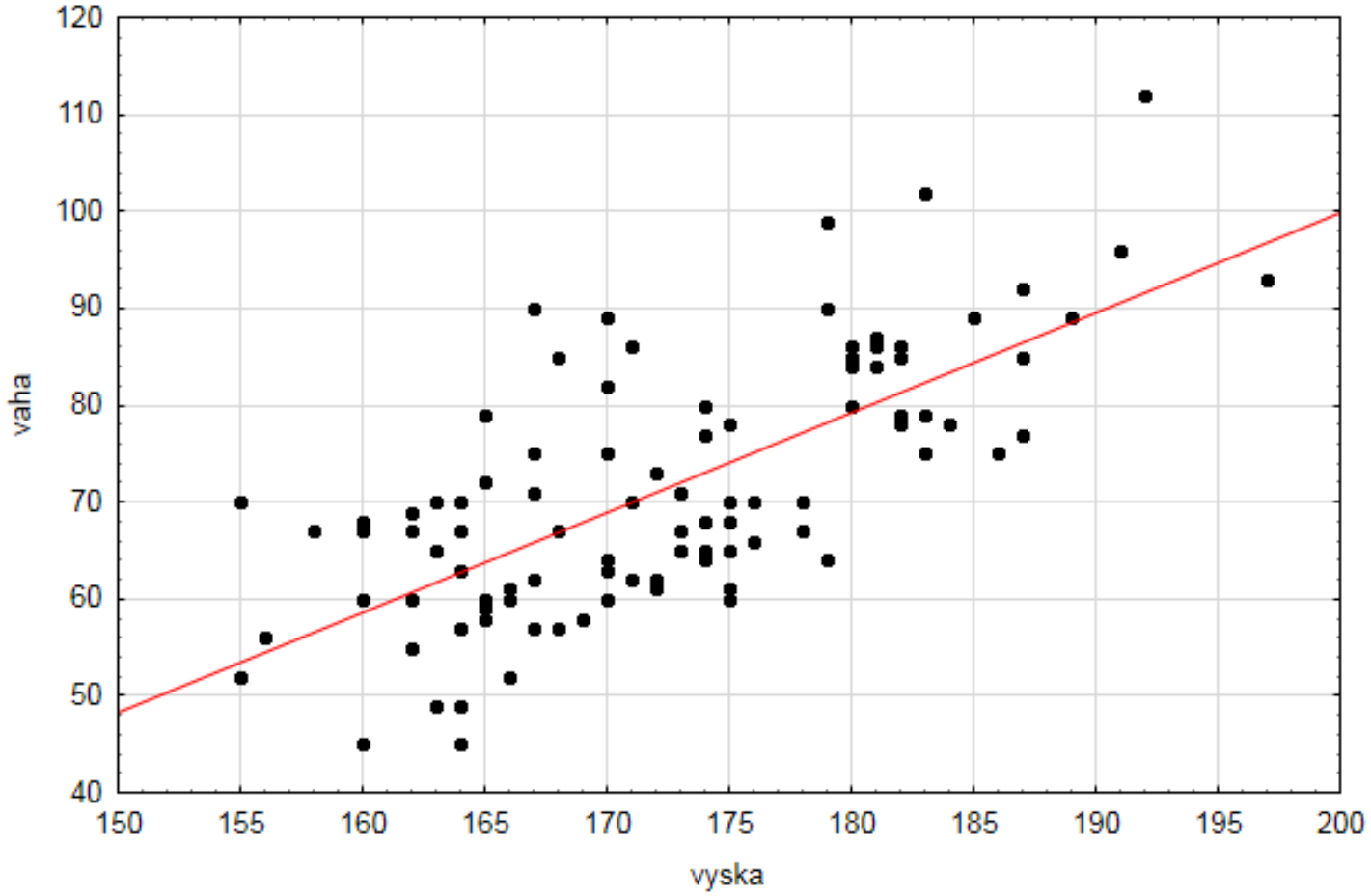
Dvourozměrný histogram

- pro vykreslení vztahu dvou kvantitativních proměnných
- v softwaru R: hist()



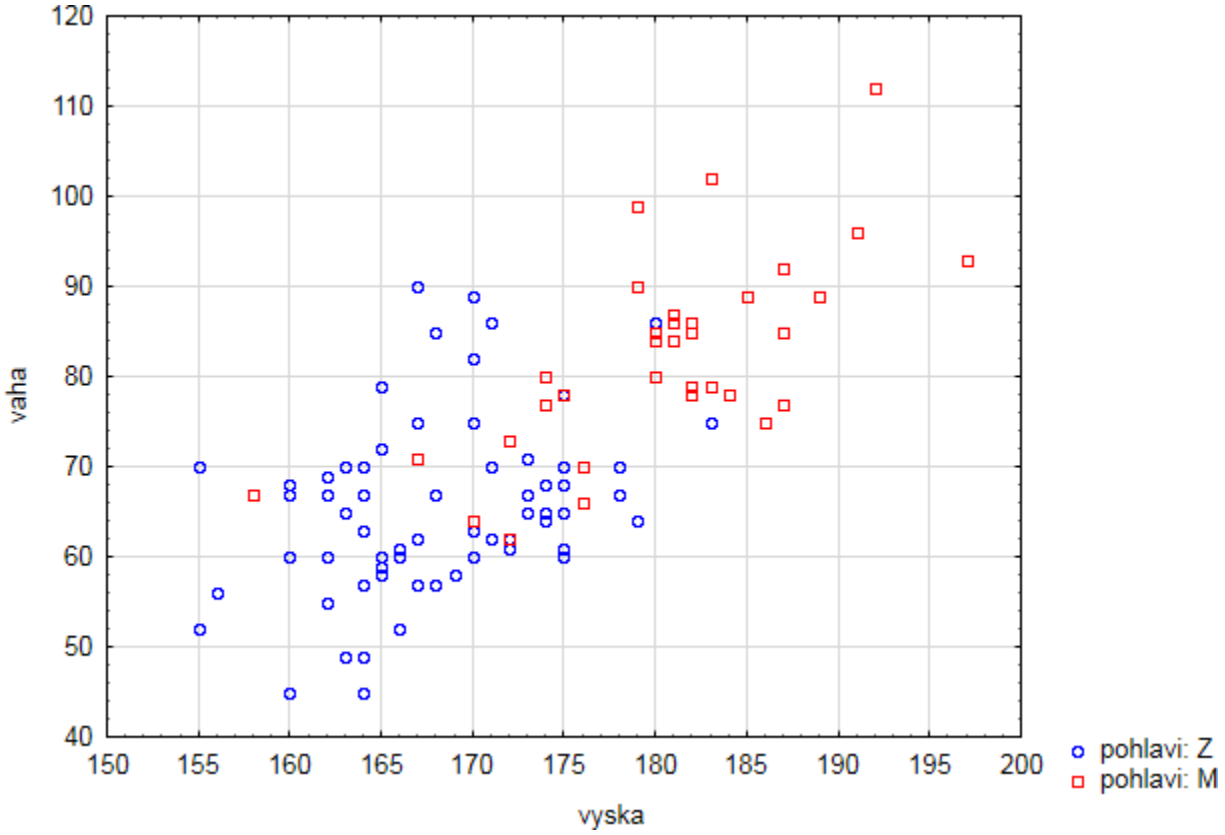
Tečkový graf

- rovněž pro vykreslení vztahu dvou kvantitativních proměnných
- v softwaru R: plot()



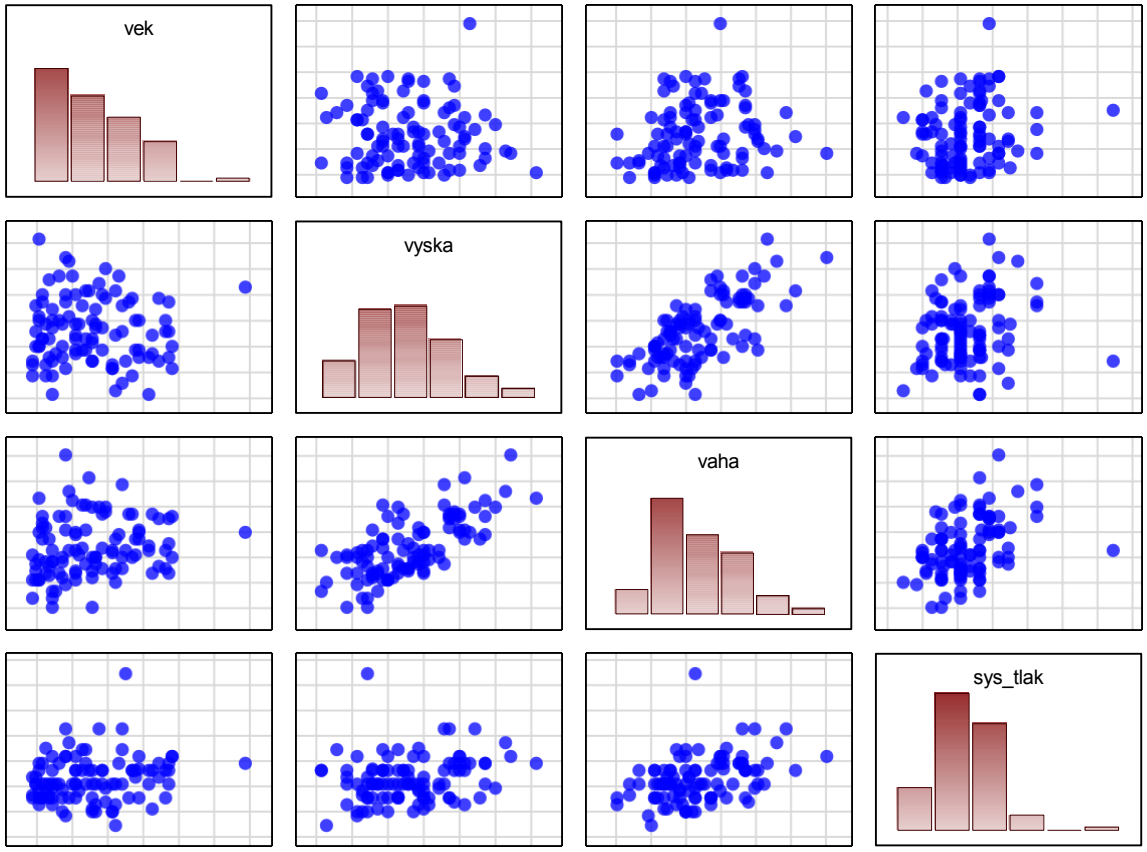
Tečkový graf – přidání kategoriální proměnné

- zahrnutí kategoriální proměnné do grafu použitím různých symbolů či barev pro jednotlivé skupiny určené danou kategoriální proměnnou



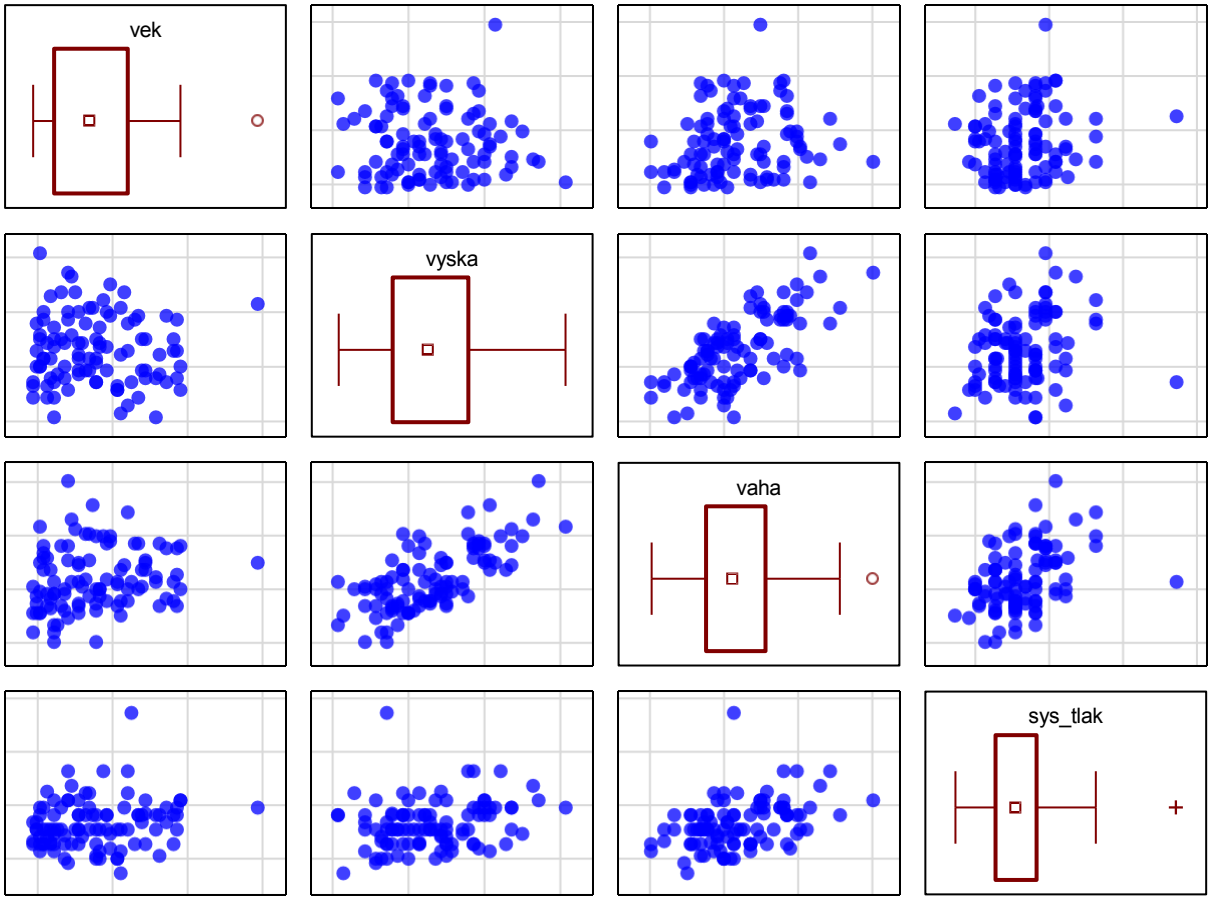
Maticový graf

- vykreslení vztahu více kvantitativních proměnných
- Na diagonále jsou histogramy
- V softwaru R: pairs(), ggpairs()



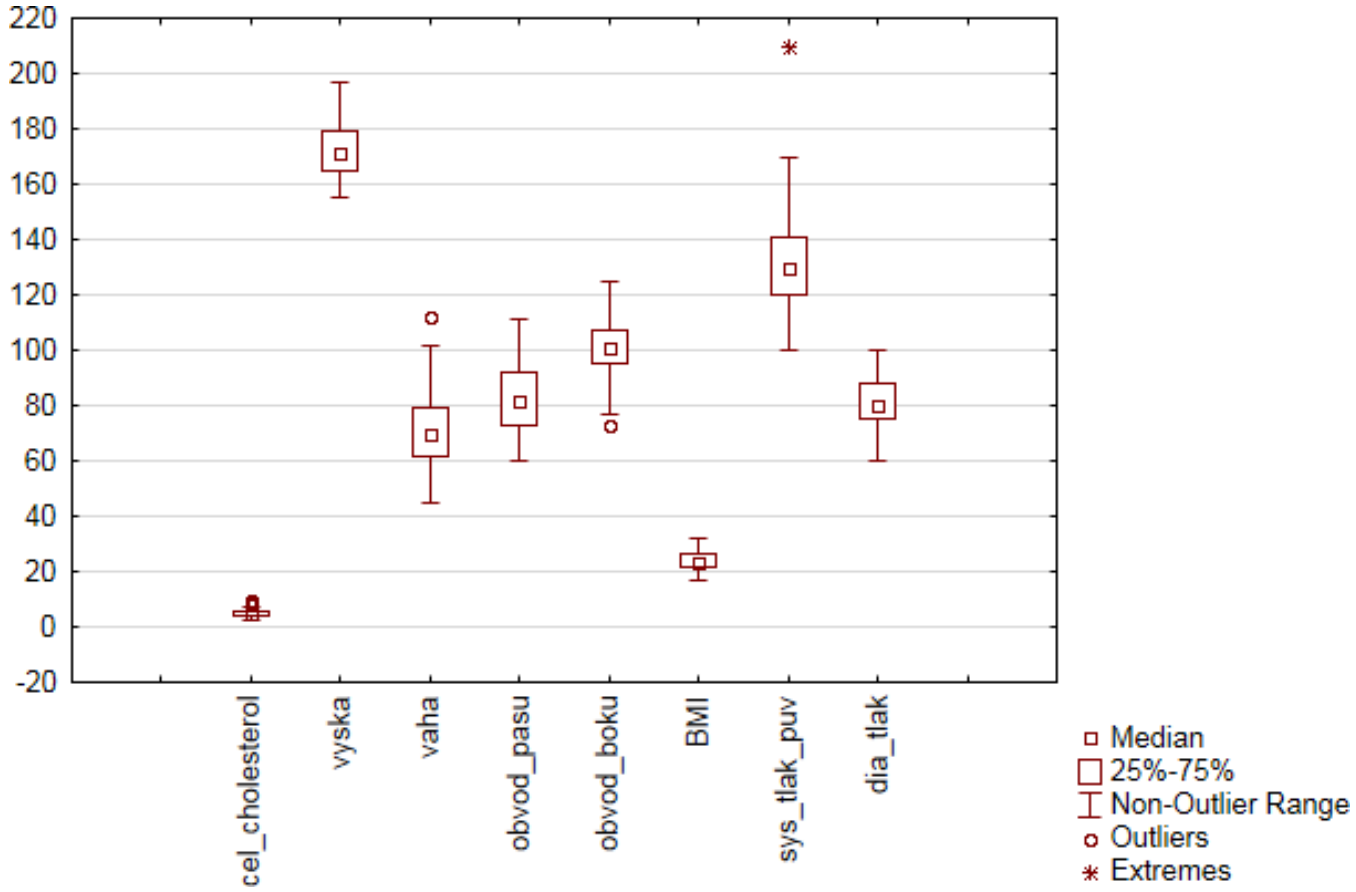
Maticový graf – na diagonále krabicové grafy

- V softwaru R: pairs(), ggpairs()



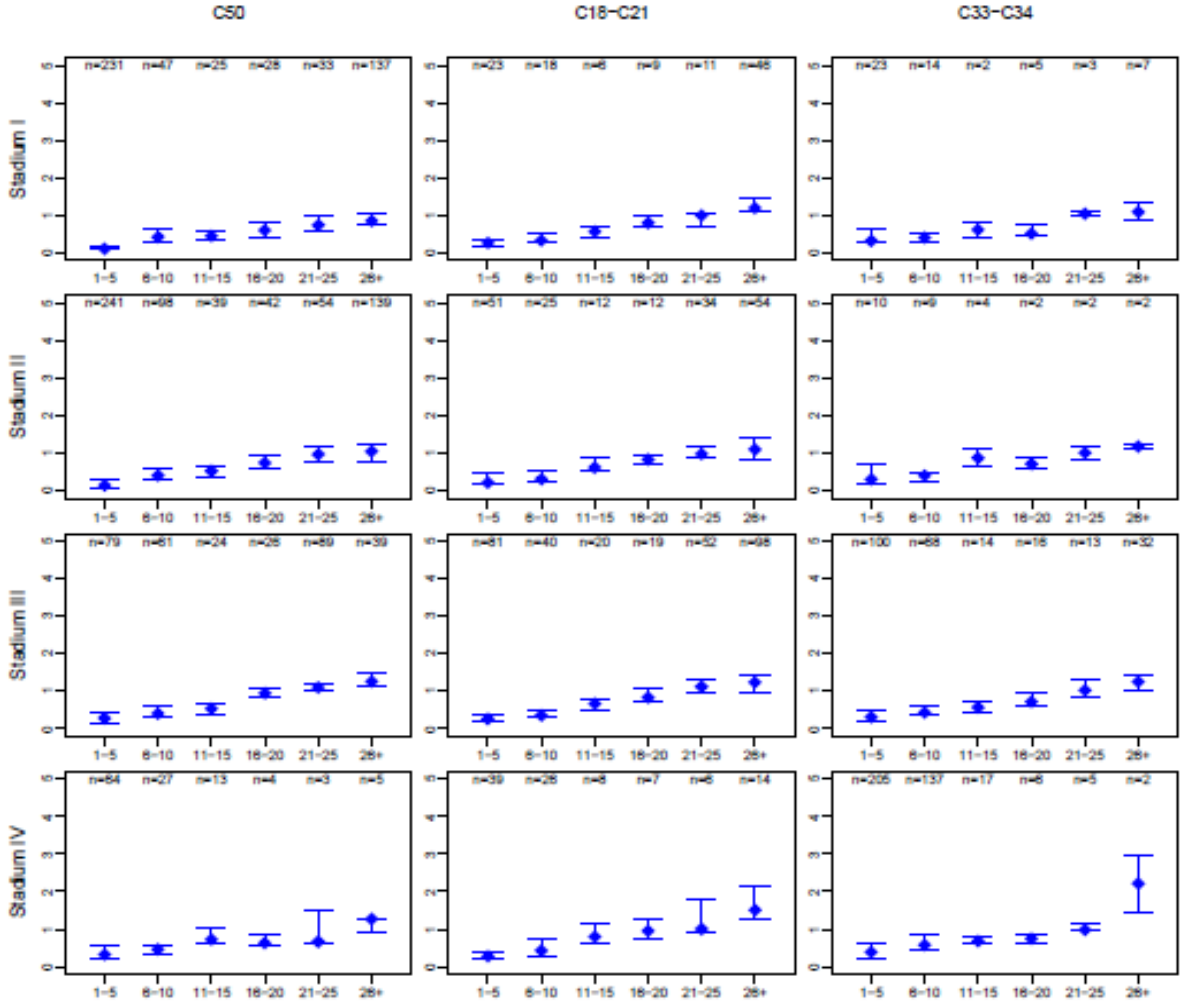
Krabicové grafy pro více proměnných

- ukáží nám, zda mají proměnné podobný rozsah hodnot
- v softwaru R: boxplot()



Vícenásobné krabicové grafy

- umožňují znázornění vztahu několika kvalitativních proměnných a jedné kvantitativní proměnné

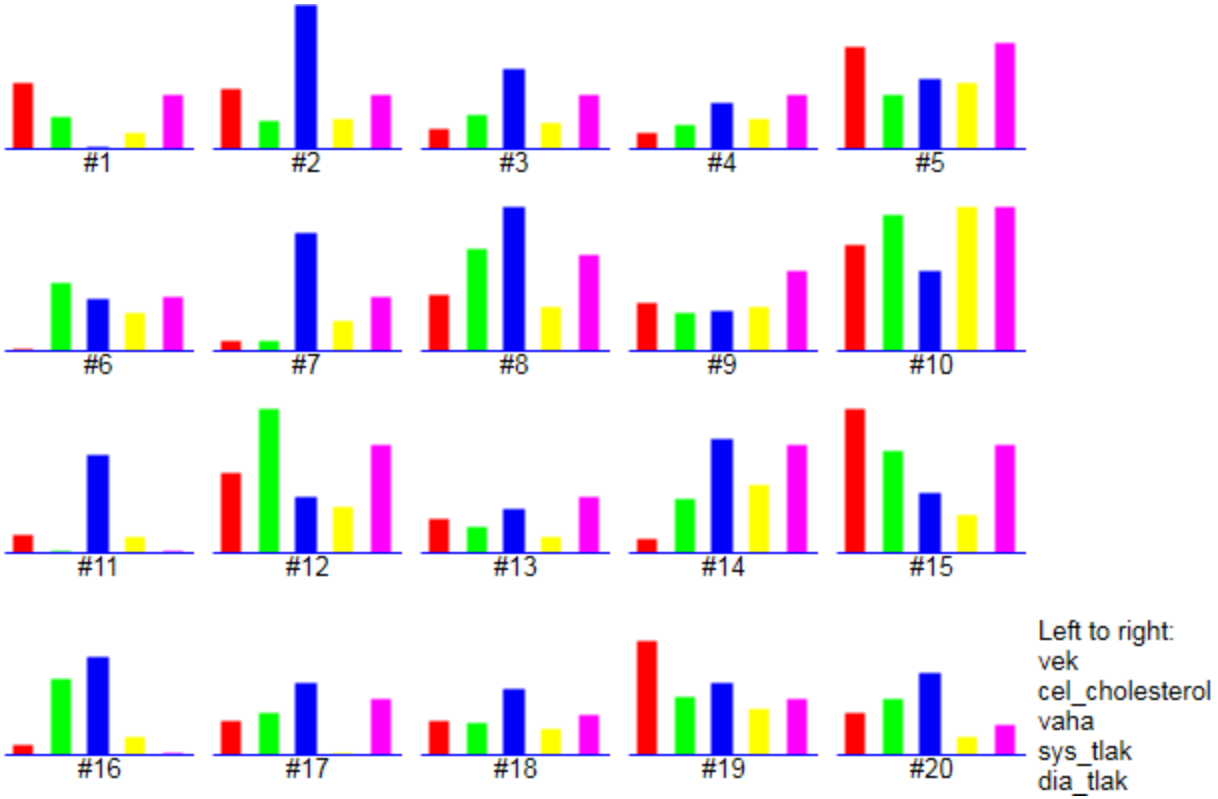


Ikonové (symbolové) grafy

- hodnoty znaků znázorněny jako geometrické útvary či symboly
- každému objektu (subjektu) odpovídá jeden obrazec složený z těchto geometrických útvarů či symbolů
- umožní vizuálně porovnat, které objekty (subjekty) jsou si podobné
- mnoho druhů, v softwaru Statistica např.:
  1. Profilové sloupce
  2. Profily
  3. Paprskové (hvězdicové) grafy
  4. Polygony
  5. Pavučinové grafy
  6. Chernoffovy tváře

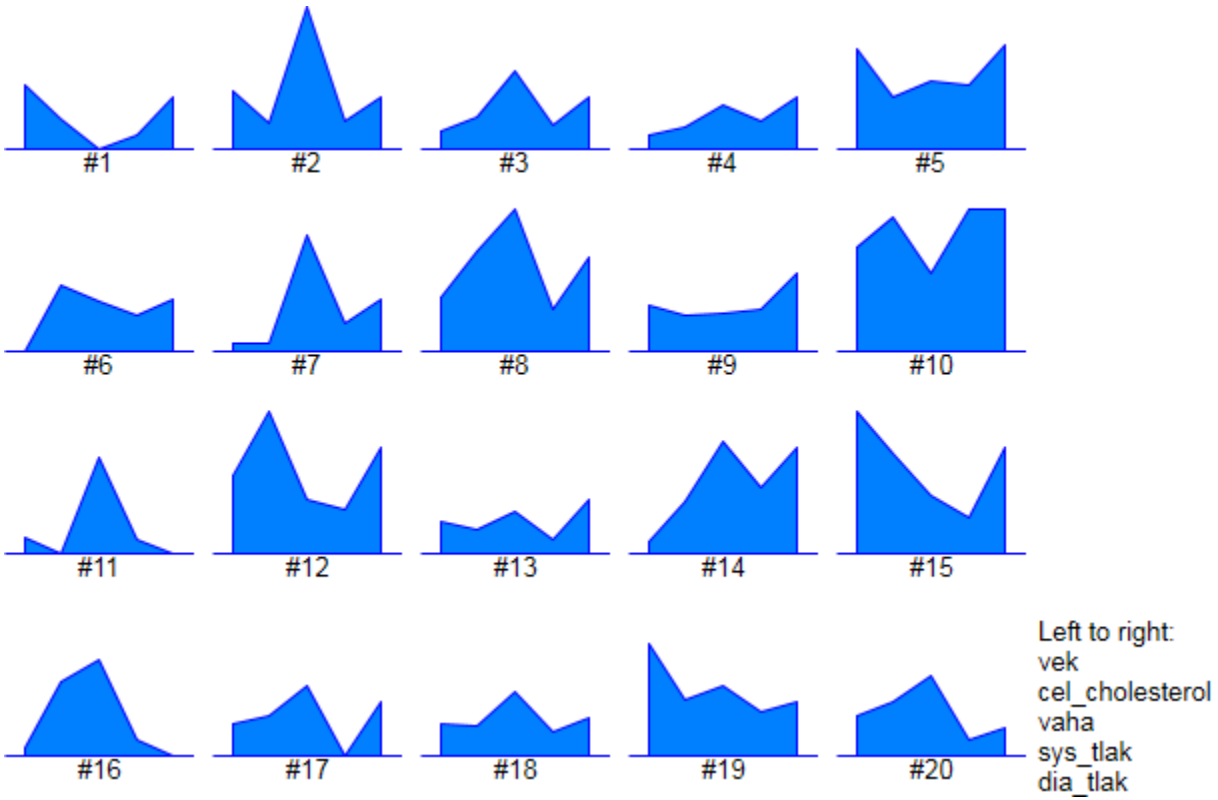
Ikonové grafy – profilové sloupce

- výšky sloupců odpovídají relativním hodnotám proměnných (relativní hodnota je podíl původní hodnoty a maxima z absolutních hodnot dané proměnné)



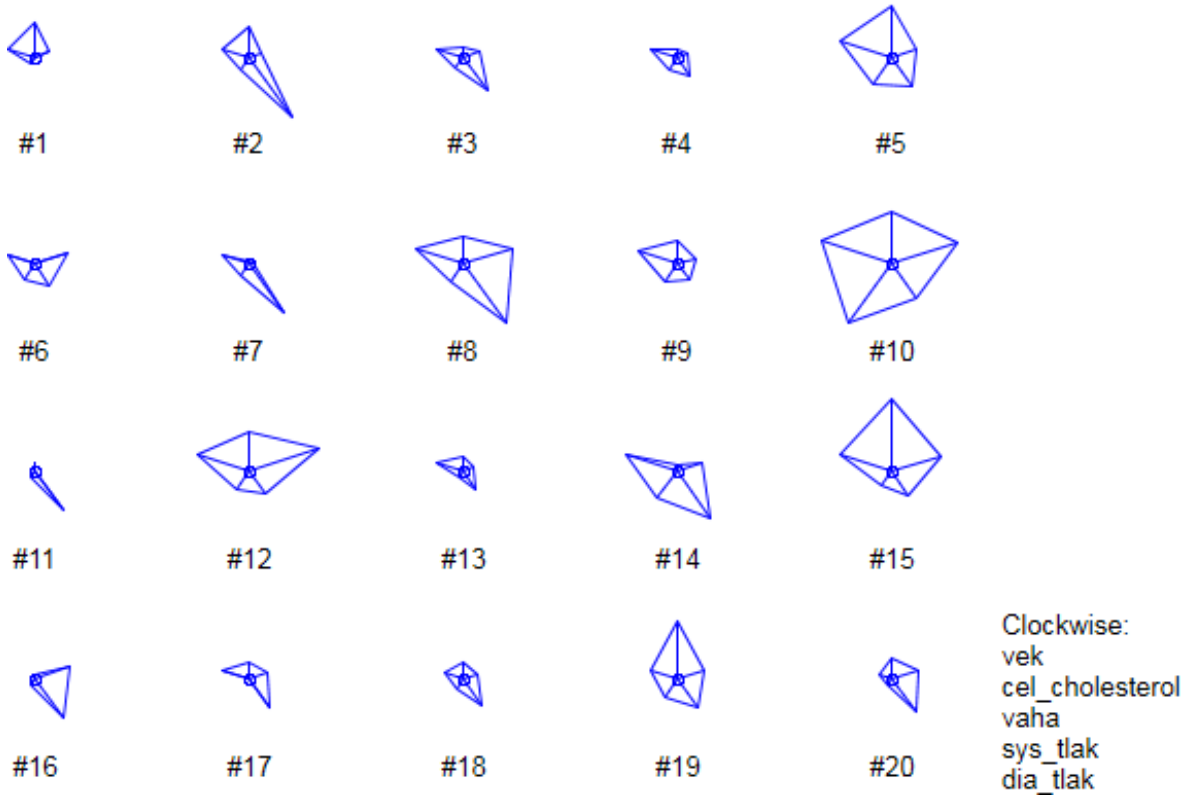
Ikonové grafy – profily

- obdoba profilových sloupců, jen se středy horních hran profilových sloupců spojí úsečkami



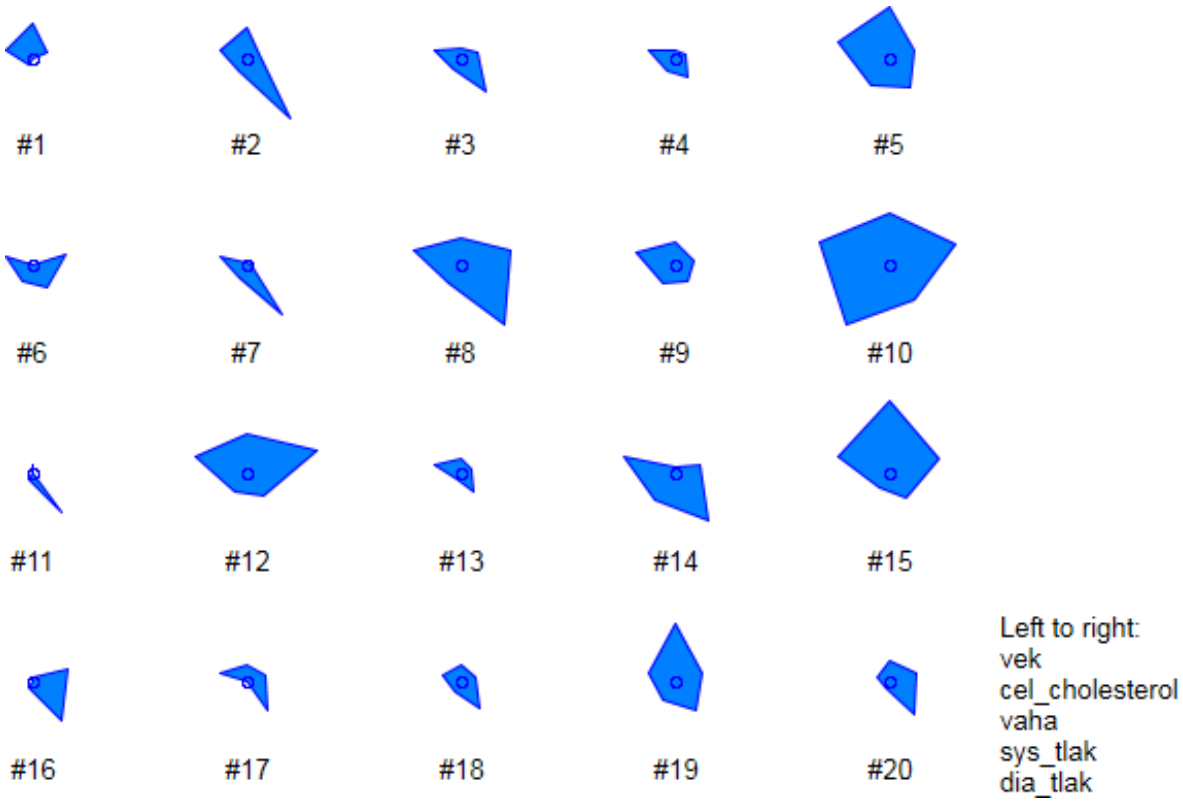
Ikonové grafy – paprskové (hvězdicové) grafy

- vzdálenosti od středu odpovídají relativním hodnotám proměnných



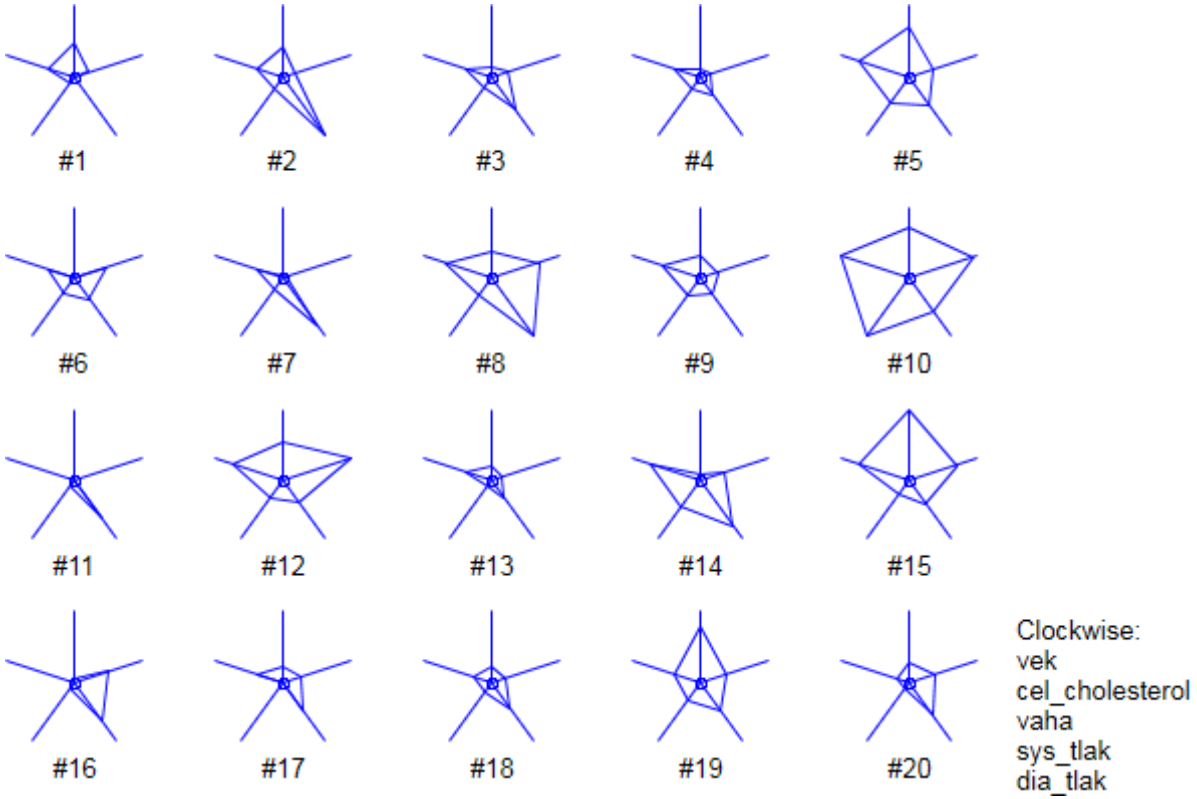
Ikonové grafy – polygony

- obdoba paprskových grafů, jen jsou vyplněné



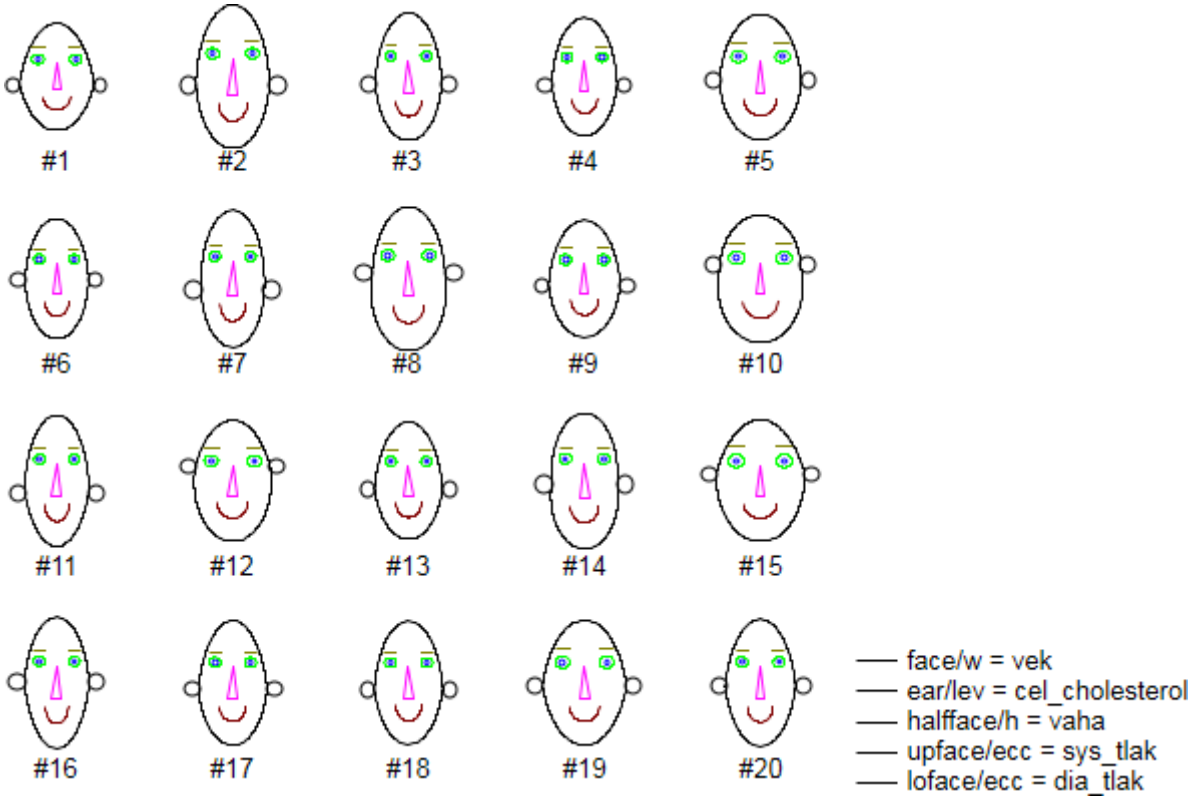
Ikonové grafy – pavučinové grafy

- obdoba paprskových grafů, přidáno znázornění maxima absolutních hodnot



Ikonové grafy – Chernoffovy tváře

- proměnné znázorněny jako části obličeje



Vizualizace vícerozměrných dat - shrnutí

- 3D sloupkové grafy
- dvourozměrný histogram
- maticové grafy
- krabicové grafy pro více proměnných
- ikonové (symbolové) grafy:
  - profilové sloupce
  - profily
  - paprskové (hvězdicové) grafy
  - polygony
  - pavučinové grafy
  - Chernoffovy tváře