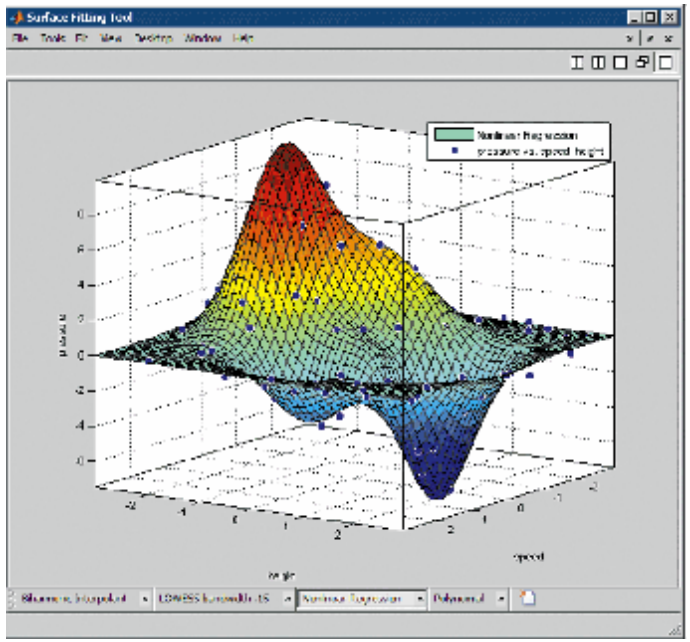
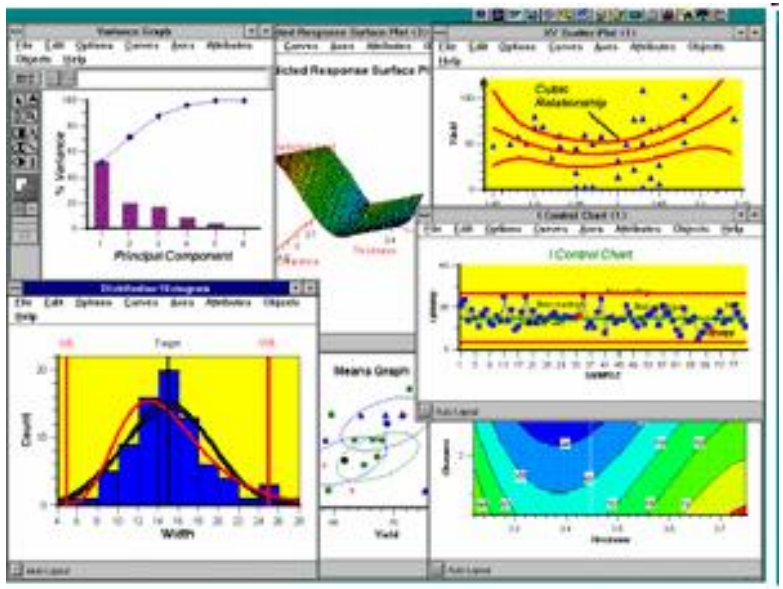
Explorační analýza, transformace dat





Exploračnı́ analýza – PROČ?

- To *listen* to the data:
  - to catch mistakes
  - to see patterns in the data
  - to find violations of statistical assumptions
  - …and because if you don’t, you will have trouble later

Čištění dat: Praktické zkušenosti

- Pokud vaše nová data obsahují více než 30 čísel, tak je v nich skoro jistě nějaká chyba
- Čištění a příprava dat zabírá obvykle víc než 80 – 90 % analytikova času
- Věnuje se mu jako hlavnímu tématu méně než 1 % článků ve statistických a podobných časopisech
- Pokud budete VELMI pečliví v této fázi, ušetříte si daleko víc času a nervů později – jinak stavíte dům na písku.

Čištění dat: Ověření souboru

- Ověření souboru s daty / zdrojů dat
  - Jsou to správná data (čas vzniku, výzkum…)?
  - Jsou kompletní, bez duplicit, umím je číst…
- Zkoumání případů
  - Mají identifikátory?
    - Jsou tyto ID správné?
  - Neopakují se (duplicity)?
    - Existují i „skoro" duplicity – dva podobné, ale ne přesně totožné záznamy o tomtéž subjektu.
  - Nejsou vynechány?

Čištění dat: Ověření proměnných

- Zkoumání metadat o proměnných
  - Jsou tam všechny proměnné a správně značené?
  - Je jasné, co znamenají (kódovníky, definice…)? Dokumentace OK?
    - Pozor na mezinárodní studie, produkty konsorcií agentur a opakované vlny výzkumů. Jemné nuance metody mohou způsobit hrubý nesoulad !
  - Neopakuje se některá proměnná vícekrát?

Čištění dat: Průzkum proměnných

- Nabývá přípustných hodnot (x out of range)?
- „Divné" kódy („xxx", „9999"…)
- Duplicitní kódy pro stejnou věc („Ž", „ž", „žena", „zena"…)
- Kódování češtiny/ruštiny/…

Čištění dat: Průzkum proměnných

- Překlepy apod.
  - Editovací distance (Levenshteinova (Владимир Иосифович Левенштейн), ...) pomohou odhalit překlep
  - Editovací distance = počet elementárních editovacích kroků potřebných pro změnu jednoho řetězce na druhý. Viz http://www.merriampark.com/ld.htm k Levenshteinově distanci
    - Je zde aplet, který ji umí počítat
  - Shlukování řetězců podle ED

Čištění dat: Průzkum proměnných

- Slučování podobných kategorií (prodavač – prodejce – prodavačka);
- Málo četné kategorie (národnost brazilská…) – co s tím?
- Je distribuce přiměřená našemu očekávání (interval hodnot, rozptyl, šikmost, špičatost, modální hodnoty…)? Není např. příliš „ořezaná“ či naopak „roztažená“?
  - Někdy se obtížně poznává: Např. věk v části dat může být kódován jako poslední dvojčíslí roku narození, a v jiné části dat jako *2007 – rok narození*

Čištění dat: Průzkum proměnných

- Shluky (clumping), typicky kolem zaokrouhlených hodnot
  - Nebo třeba kolem hranic věkových kvót, vzniklé tím, jak tazatelé „upravují“ věky respondentů, aby se vešli do kvót
- Chybějící hodnoty (příčiny vzniku, zastoupení,…)!!!
- Pozor na kódy časů (amer. x evrop. konvence), regionů apod.!

Čištění dat: Vazby mezi daty

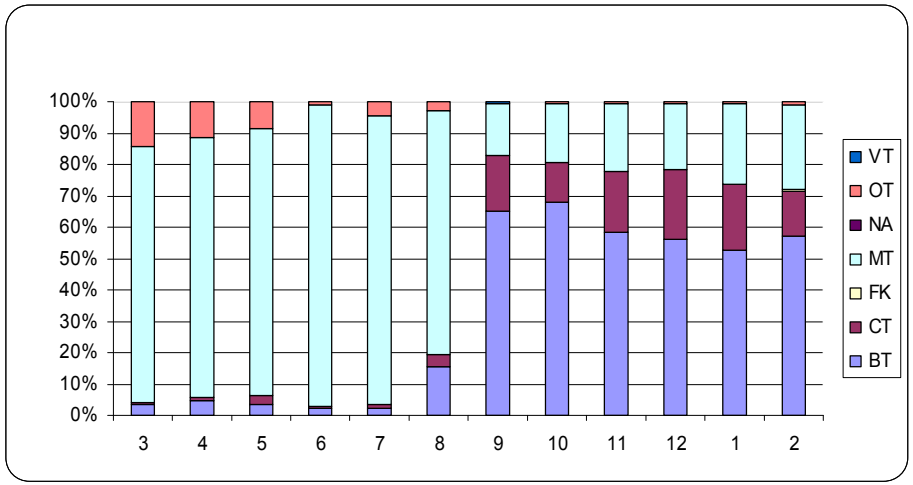
- Více proměnných
  - Kontingenční tabulky, box ploty s kategoriemi, bodové grafy a jejich matice, korelační koeficienty
  - Logické vazby (např. 10tiletý nemůže být ženatý, 30tiletý nemůže pracovat 20let,…)
    - Hledání pomocí programu – podmínky vyjádříme pomocí prostředků matematické logiky a necháme počítač, aby vyhledal případy, kde nejsou splněny.

Čištění dat: Vazby mezi daty

- Více proměnných
  - Extrémní hodnoty vícerozměrného rozdělení
    - Bodový graf
    - Mahalanobisova vzdálenost od těžiště: $[(\boldsymbol{x}-\boldsymbol{t})^T \boldsymbol{S}^{-1} (\boldsymbol{x}-\boldsymbol{t})]^{-1/2}$, kde $\boldsymbol{t}$ je vektor těžiště, $\boldsymbol{x}$ zkoumaný bod a $\boldsymbol{S}$ kovarianční matice
      - např. P. Filzmoser (2004) A multivariate outlier detection method, http://www.statistik.tuwien.ac.at/public/filz/papers/minsk04.pdf
  - Další vlastnosti; např. existují očekávané korelace?

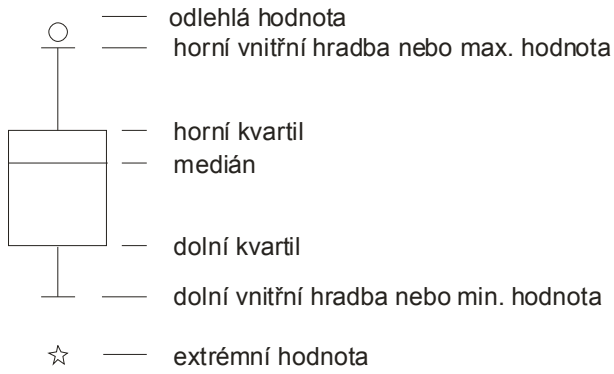
Čištění dat: Vazby mezi daty

- korektní vkládání dat do DB
  - text. pole s názvem zboží vs. rolovací seznam s typem zboží



  - pořadí hodnot v rolovacím seznamu – problém první (defaultní) hodnoty

Čištění dat: Odlehlé hodnoty



- kvartilová odchylka: $q = x_{0.75} - x_{0.25}$
- vnitřní hradby: $x_{0.25} - 1.5q$, $x_{0.75} + 1.5q$
- vnější hradby: $x_{0,25} - 3q$, $x_{0,75} + 3q$

- **Odlehlá hodnota** leží mezi vnějšími a vnitřními hradbami, tj. v intervalu $(x_{0,75} + 1,5q, x_{0,75} + 3q)$ či v intervalu $(x_{0,25} - 3q, x_{0,25} - 1,5q)$.
- **Extrémní hodnota** leží za vnějšími hradbami, tj. v intervalu $(x_{0,75} + 3q, \infty)$
- či v intervalu $(-\infty, x_{0,25} - 3q)$.

Čištění dat: Opravy chyb

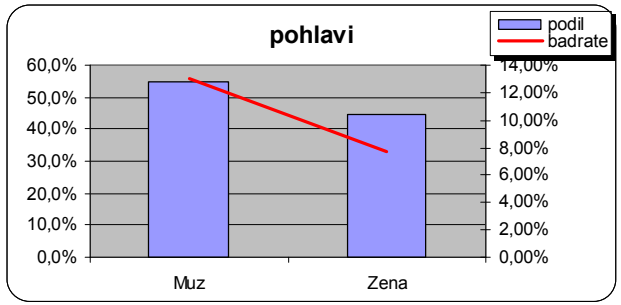
- Zpět k pramenům!
- Vyřazení podezřelých případů
  - Záměrné podvody, např. nespolehliví tazatelé (shluková analýza!)
  - Neověřitelná data
- Vyřazení podezřelých hodnot
- Rekódování na správné hodnoty (imputace hodnot)
  - imputace – průměrem, mediánem, max./min. hodnotou, pomocí modelu

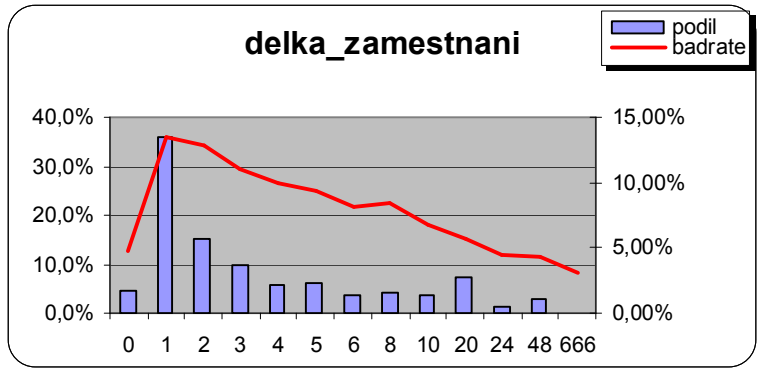
Explorace dat - jednorozměrná

- Frekvenční tabulky, histogramy:

	pocet	podil	badrate
Muz	248 768	55,0%	13,08%
Zena	203 194	45,0%	7,69%
Total	451 962	100,0%	10,66%

delka_zamestnani	pocet	podil	badrate
0	20 825	4,6%	4,69%
1	163 144	36,1%	13,43%
2	67 462	14,9%	12,80%
3	43 778	9,7%	10,97%
4	26 256	5,8%	10,01%
5	27 526	6,1%	9,32%
6	15 893	3,5%	8,16%
8	18 036	4,0%	8,39%
10	17 195	3,8%	6,72%
20	33 641	7,4%	5,60%
24	5 176	1,1%	4,48%
48	12 934	2,9%	4,28%
666	96	0,0%	3,13%
Total	451 962	100,0%	10,66%

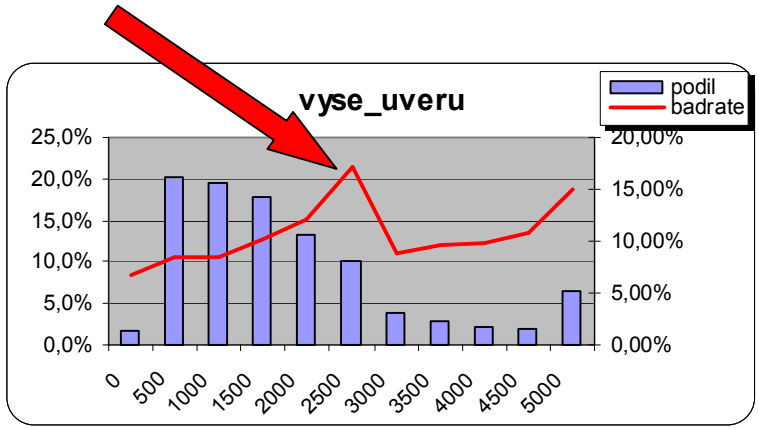




Explorace dat - jednorozměrná

- výše úvěru vs. bad rate

OK? Nebo je to způsobeno jiným faktorem???

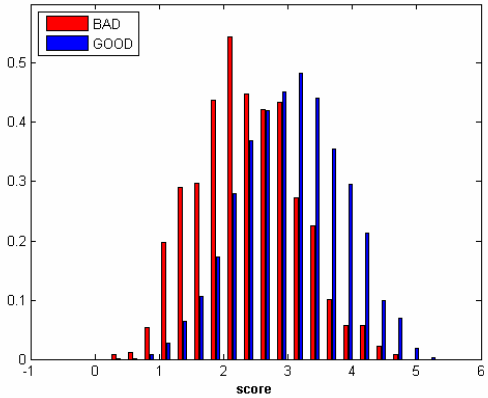


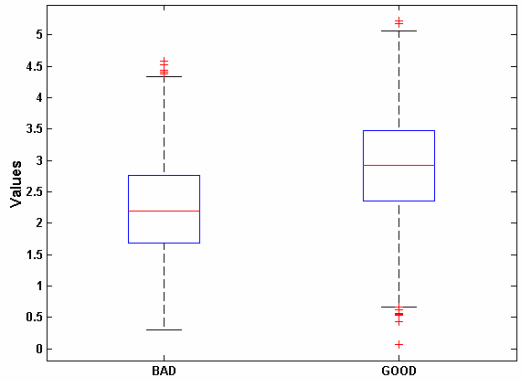
Explorace dat - jednorozměrná

- spojité proměnné:
  - Průměr
  - Modus
  - Kvantily
  - Rozptyl
  - Min./maximální hodnota
- je vhodná kategorizace

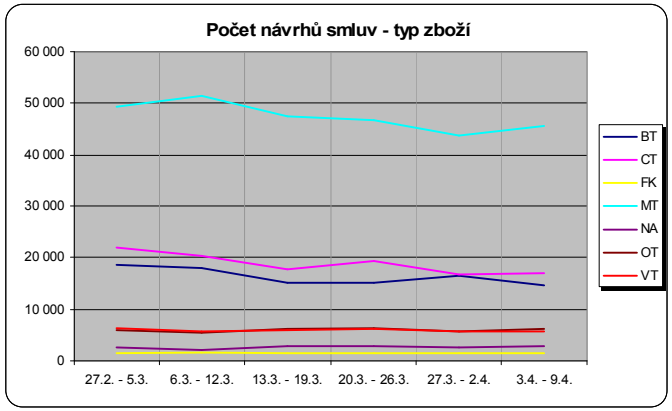
Explorace dat - jednorozměrná

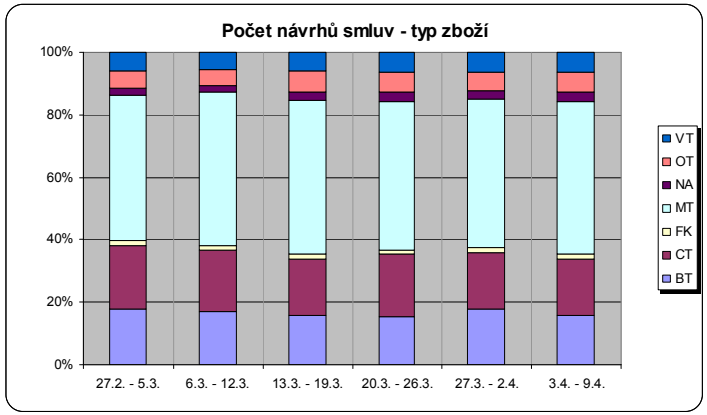
- Histogramy, box ploty





- Stabilita v čase





Explorace dat - vícerozměrná

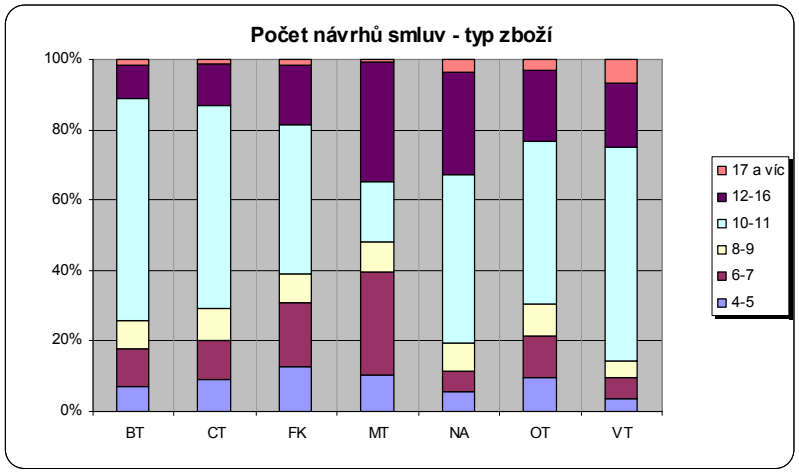
- Kontingenční tabulky

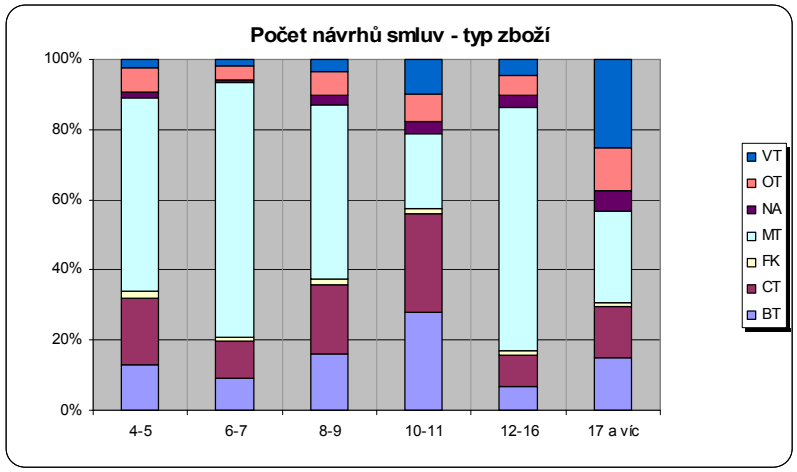
	do 5 000	5 000 - 10 000	10 000 - 15 000	víc než 15 000
BT	4 291	8 581	9 176	9 044
CT	7 587	12 493	6 500	7 236
FK	258	1 017	851	557
MT	27 191	39 551	16 524	5 992
NA	426	1 088	1 114	2 737
OT	2 478	3 689	2 103	3 475
VT	384	1 001	963	9 086

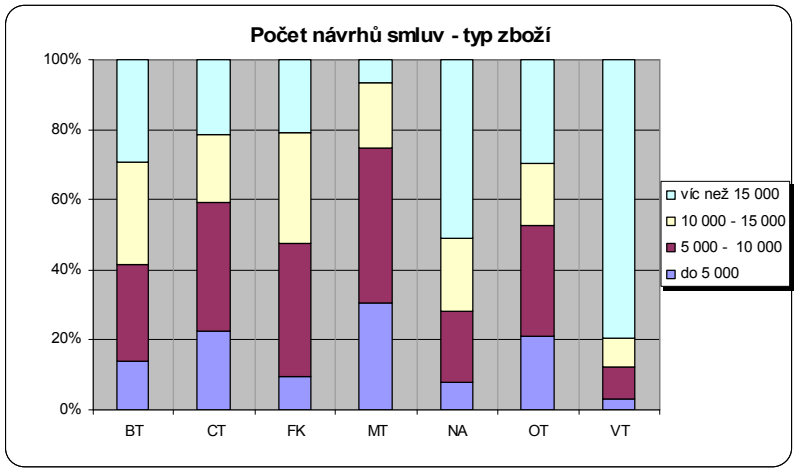
row%	do 5 000	5 000 - 10 000	10 000 - 15 000	víc než 15 000
BT	13,8%	27,6%	29,5%	29,1%
CT	22,4%	36,9%	19,2%	21,4%
FK	9,6%	37,9%	31,7%	20,8%
MT	30,5%	44,3%	18,5%	6,7%
NA	7,9%	20,3%	20,8%	51,0%
OT	21,1%	31,4%	17,9%	29,6%
VT	3,4%	8,8%	8,4%	79,5%

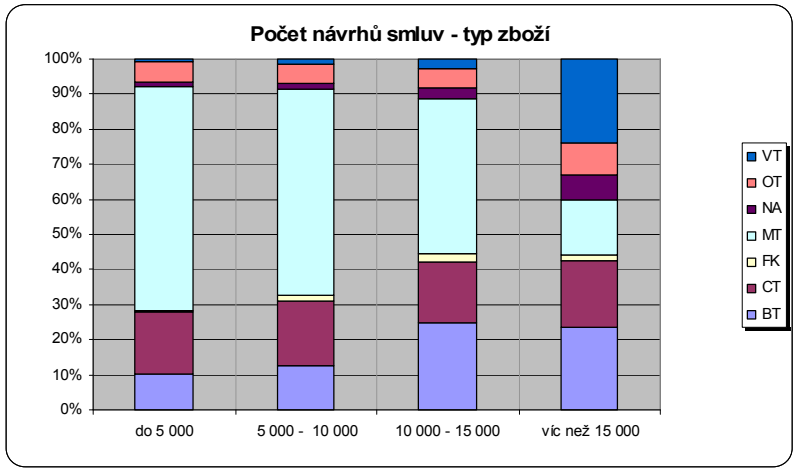
col%	do 5 000	5 000 - 10 000	10 000 - 15 000	víc než 15 000
BT	10,1%	12,7%	24,6%	23,7%
CT	17,8%	18,5%	17,5%	19,0%
FK	0,6%	1,5%	2,3%	1,5%
MT	63,8%	58,7%	44,4%	15,7%
NA	1,0%	1,6%	3,0%	7,2%
OT	5,8%	5,5%	5,6%	9,1%
VT	0,9%	1,5%	2,6%	23,8%

Explorace dat - vícerozměrná



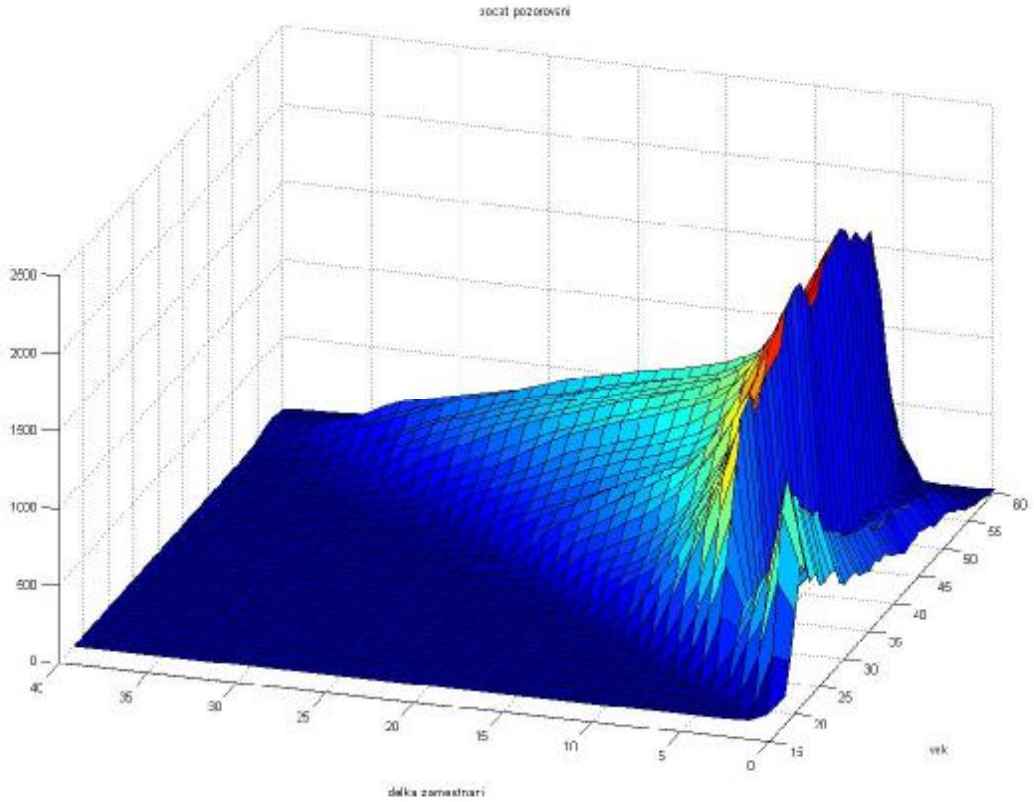






Explorace dat - vícerozměrná

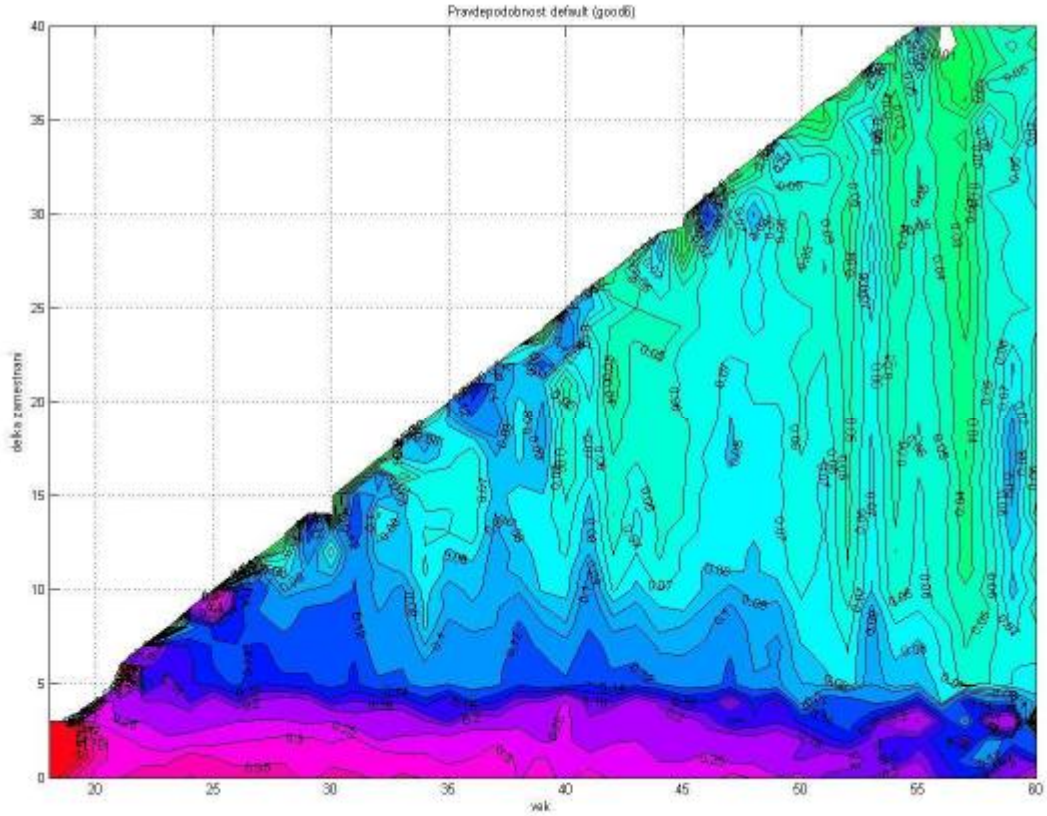
- Věk vs. délka zaměstnání



5 let
…defaultní
hodnota???

Explorace dat - vícerozměrná

- Věk vs. délka zaměstnání vs. default



Diskriminační síla proměnných

- IV (Information Value)

$$I_{val} = \int_{-\infty}^{\infty} \left(f_{GOOD}(x) - f_{BAD}(x)\right) \ln\left(\frac{f_{GOOD}(x)}{f_{BAD}(x)}\right) dx$$

score int.	# bad clients	#good clients	% bad [1]	% good [2]	[3] = [2] - [1]	[4] = [2] / [1]	[5] = ln[4]	[6] = [3] * [5]
1	1	10	2,0%	1,1%	-0,01	0,53	-0,64	0,01
2	2	15	4,0%	1,6%	-0,02	0,39	-0,93	0,02
3	8	52	16,0%	5,5%	-0,11	0,34	-1,07	0,11
4	14	93	28,0%	9,8%	-0,18	0,35	-1,05	0,19
5	10	146	20,0%	15,4%	-0,05	0,77	-0,26	0,01
6	6	247	12,0%	26,0%	0,14	2,17	0,77	0,11
7	4	137	8,0%	14,4%	0,06	1,80	0,59	0,04
8	3	105	6,0%	11,1%	0,05	1,84	0,61	0,03
9	1	97	2,0%	10,2%	0,08	5,11	1,63	0,13
10	1	48	2,0%	5,1%	0,03	2,53	0,93	0,03
All	50	950					Info. Value	0,68

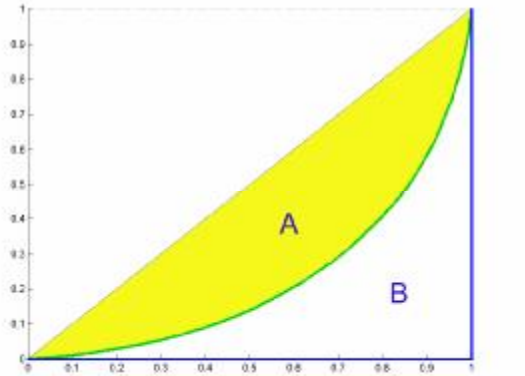
Diskriminační síla proměnných

- IV (Information Value)
  - <0.02 unpredictive
  - 0.02 – 0.1 weak
  - 0.1 – 0.3 medium
  - 0.3 – 0.5 strong
  - > 0.5 too high …je třeba prověřit, pravděpodbně je něco špatně

Diskriminační síla proměnných

- Lorenzova křivka, Giniho index

$$x = F_{m.BAD}(a)$$
$$y = F_{n.GOOD}(a),\ a \in [L, H].$$

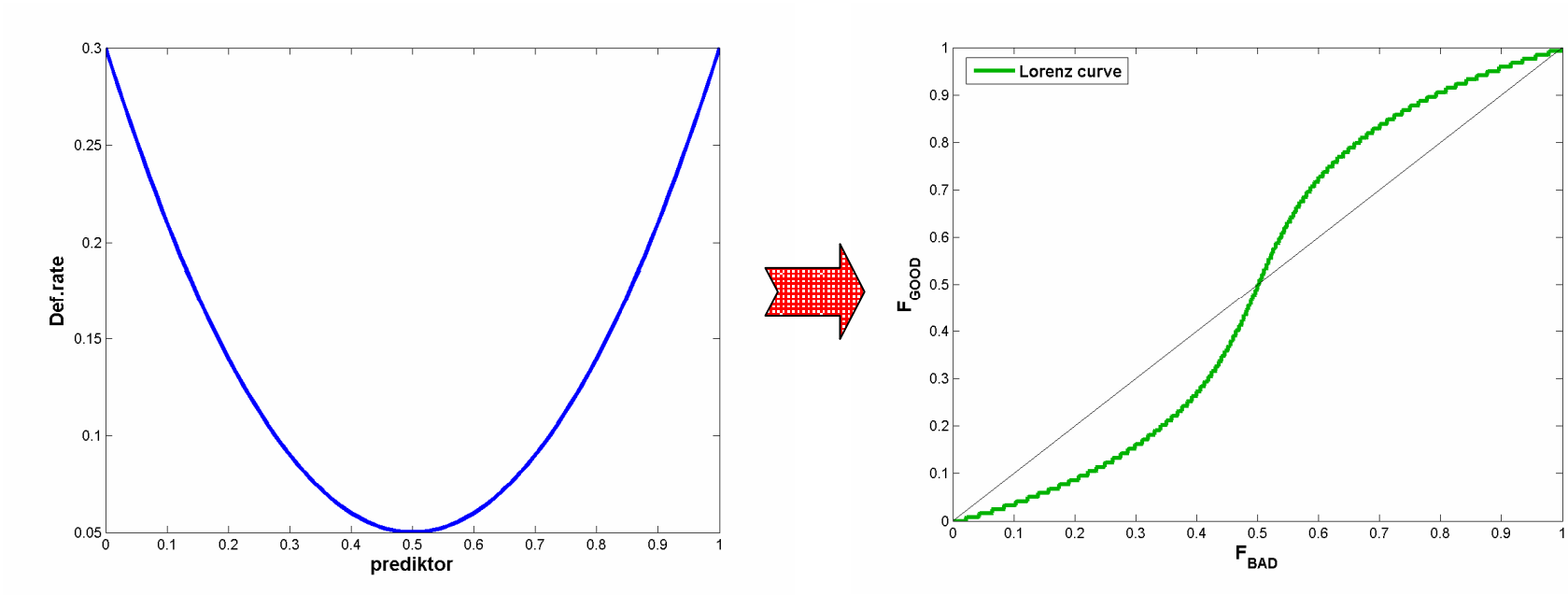


$$Gini = \frac{A}{A+B} = 2A$$

$$Gini = 1 - \sum_{k=2}^{n+m} \left(F_{m.BAD\,k} - F_{m.BAD\,k-1}\right) \cdot \left(F_{n.GOOD_{\,k}} + F_{n.GOOD\,k-1}\right)$$

Diskriminační síla proměnných

- Lorenzova křivka …kontrola monotónnosti PD na dané proměnné



Kategorizace (WOE)

Diskriminační síla proměnných

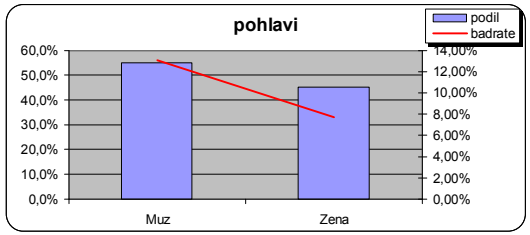
- Giniho index

  - <0.05 unpredictive
  - 0.05 – 0.1 weak
  - 0.1 – 0.2 medium
  - 0.2 – 0.5 strong
  - > 0.5 too high …je třeba prověřit, pravděpodbně je něco špatně

Diskriminační síla proměnných

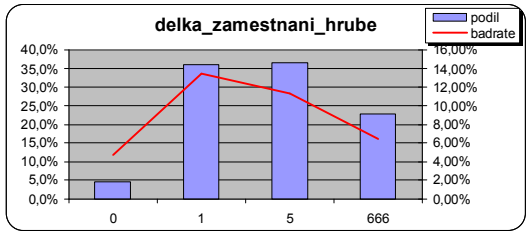
pohlavi		Gini:	0,1401	Info.Value:	0,0828

	pocet	podil	badrate
Muz	248 768	55,0%	13,08%
Zena	203 194	45,0%	7,69%
Total	451 962	100,0%	10,66%



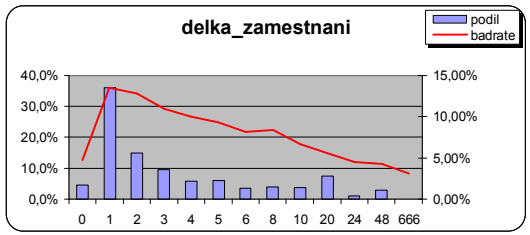
delka_zamestnani_hrube		Gini:	0,1611	Info.Value:	0,1100

	pocet	podil	badrate
0	20 825	4,6%	4,69%
1	163 144	36,1%	13,43%
5	165 022	36,5%	11,29%
666	102 971	22,8%	6,45%
Total	451 962	100,0%	10,66%



delka_zamestnani_jemne		Gini:	0,1762	Info.Value:	0,1285

delka_zamestnani	pocet	podil	badrate
0	20 825	4,6%	4,69%
1	163 144	36,1%	13,43%
2	67 462	14,9%	12,80%
3	43 778	9,7%	10,97%
4	26 256	5,8%	10,01%
5	27 526	6,1%	9,32%
6	15 893	3,5%	8,16%
8	18 036	4,0%	8,39%
10	17 195	3,8%	6,72%
20	33 641	7,4%	5,60%
24	5 176	1,1%	4,48%
48	12 934	2,9%	4,28%
666	96	0,0%	3,13%
Total	451 962	100,0%	10,66%



Transformace dat

- binarizace (dummy proměnné)
  - Dummy variables refers to the technique of using a dichotomous variable (coded 0 or 1) to represent the separate categories of a nominal level measure.
  - The term “dummy” appears to refer to the fact that the presence of the trait indicated by the code of 1 represents a factor or collection of factors that are not measurable by any better means within the context of the analysis.

Dummy Variables

- Dummy variable involves assigning 1 to observation of the chosen characteristic and 0 for the rest.
- For gender (2 categories), assign 1 for female observation and 0 for male. Only one dummy variable is created.
- For race(4 categories), we need to create more than 1 variable.
  - 1st variable=1 if African and 0 fo all other races.
  - 2nd variable=1 if White and 0 for all other races.
  - 3rd variable=1 if Asian and 0 for all other races
  - 4rth variable=1 if colored and 0 for all other races
- Important: All 4 variables are not included in regression analysis (causes perfect multicollinearity D4=1-D3-D2-D1)
- No. of dummy variables=no. of categories -1
- Omitted variable is the benchmark variable.
- Constant indicates the omitted benchmark variable
- Coefficients of the included variable is considered in relation to the constant

Transformace dat

- Kategorizace spojitých proměnných
  - decily
- Agregace
- Segmentace

Transformace dat - WOE

- **Good** celkový počet dobrých klientů ve vzorku
- **Bad** celkový počet špatných klientů ve vzorku

- **$good_i^s$, bad_i^s** počet dobrých, resp. špatných klientů v *i*-té kategorii příslušné *s-té* proměnné.
- celková šance $$odds\_all = \frac{good}{bad}$$
- šance i-té kategorie s-té proměnné $$odds_i^{\ s} = \frac{good_i^{\ s}}{bad_i^{\ s}}$$
- poměr šancí (OR) $$odds\_ratio_i^{\ s} = \frac{odds_i^{\ s}}{odds\_all}$$
- WOE (weights of evidence)

$$WOE_i^{\ s} = \ln\left(odds\_ratio_i^{\ s}\right) = \ln\left(\frac{\frac{good_i^{\ s}}{bad_i^{\ s}}}{\frac{good}{bad}}\right) = \ln\left(\frac{\frac{good_i^{\ s}}{good}}{\frac{bad_i^{\ s}}{bad}}\right)$$

Transformace dat -WOE

cat.	# bad clients	#good clients	Def_rate	odds	OR	% bad [1]	% good [2]	[3] = [2] / [1]	WOE = ln[3]
1	4	1	80,0%	0,25	0,03	40,0%	1,1%	0,03	-3,58
2	2	6	25,0%	3,00	0,33	20,0%	6,7%	0,33	-1,10
3	2	18	10,0%	9,00	1,00	20,0%	20,0%	1,00	0,00
4	1	12	7,7%	12,00	1,33	10,0%	13,3%	1,33	0,29
5	1	53	1,9%	53,00	5,89	10,0%	58,9%	5,89	1,77
All	10	90	10,0%	9,00					
ALL	100								

80% = 4/ (4+1)
0,25 = 1/4
0,03 = 0,25 / 9
40% = 4 / 10
1,1% = 1 / 90