

Tato prezentace je součástí wiki-prezentace [Metoda GUHA, LISp-Miner a typové úlohy](#)

Je dostupná z [této adresy](#)

Verse 20. 8. 2019

Typ úlohy: Pozitivní ordinální asociace

Data: [Hotel](#)

Problém: *Existuje kombinace hodnot atributů z Pobyť, Host, Bydliště, Meteo tak, že pro tuto kombinaci je zřetelná pozitivní ordinální závislost mezi některou dvojicí atributů z DHodnoceni, DPersonal, DStrava, DUbyťování a DZabava?*

Jan Rauch

Katedra informačního a znalostního inženýrství

Vysoká škola ekonomická v Praze

© Jan Rauch

# Kontingenční tabulka – příklad 1

Matice dat: [HotelPlusExterni](#)

KL( Hvek, PCenaCelkem, [HotelPlusExterni](#))

|      |             | PCenaCelkem |       |        |       |          |
|------|-------------|-------------|-------|--------|-------|----------|
|      |             | nejnižší    | nižší | průměr | vyšší | nejvyšší |
| Hvek | pod 21      | 22          | 30    | 26     | 25    | 17       |
|      | od 21 do 28 | 37          | 53    | 46     | 50    | 44       |
|      | od 28 do 60 | 199         | 194   | 204    | 207   | 216      |
|      | 60 a více   | 140         | 118   | 111    | 131   | 130      |

$\Sigma = 2000$

# Kontingenční tabulka – příklad 2

Matice dat: HotelPlusExterni / HStat(ČR)

KL( Hvek, PCenaCelkem, HotelPlusExterni / HStat(ČR) )

|      |             | PCenaCelkem |       |        |       |          |
|------|-------------|-------------|-------|--------|-------|----------|
|      |             | nejnižší    | nižší | průměr | vyšší | nejvyšší |
| Hvek | pod 21      | 16          | 16    | 11     | 11    | 10       |
|      | od 21 do 28 | 24          | 28    | 17     | 15    | 22       |
|      | od 28 do 60 | 98          | 89    | 69     | 74    | 106      |
|      | 60 a více   | 74          | 57    | 45     | 61    | 70       |

$\Sigma = 913$

# KL-tabulka

Matice dat:  $\mathcal{M}/\chi$

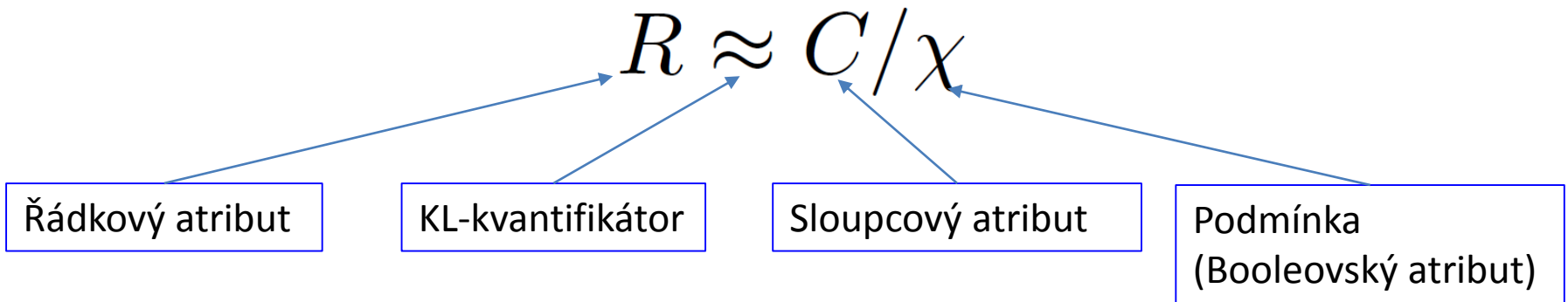
$KL(R, C, \mathcal{M}/\chi)$

|     |            | $C$                |         |           |           |            |
|-----|------------|--------------------|---------|-----------|-----------|------------|
|     |            | $\mathcal{M}/\chi$ | $c_1$   | $\dots$   | $c_L$     | $\Sigma_l$ |
| $R$ | $r_1$      | $n_{1,1}$          | $\dots$ | $n_{1,L}$ | $n_{1,*}$ |            |
|     | $\vdots$   | $\vdots$           |         | $\vdots$  | $\vdots$  |            |
|     | $r_K$      | $n_{K,1}$          | $\dots$ | $n_{K,L}$ | $n_{K,*}$ |            |
|     | $\Sigma_k$ | $n_{*,1}$          | $\dots$ | $n_{*,L}$ | $n$       |            |

$\Sigma = n$

# KL - vztah

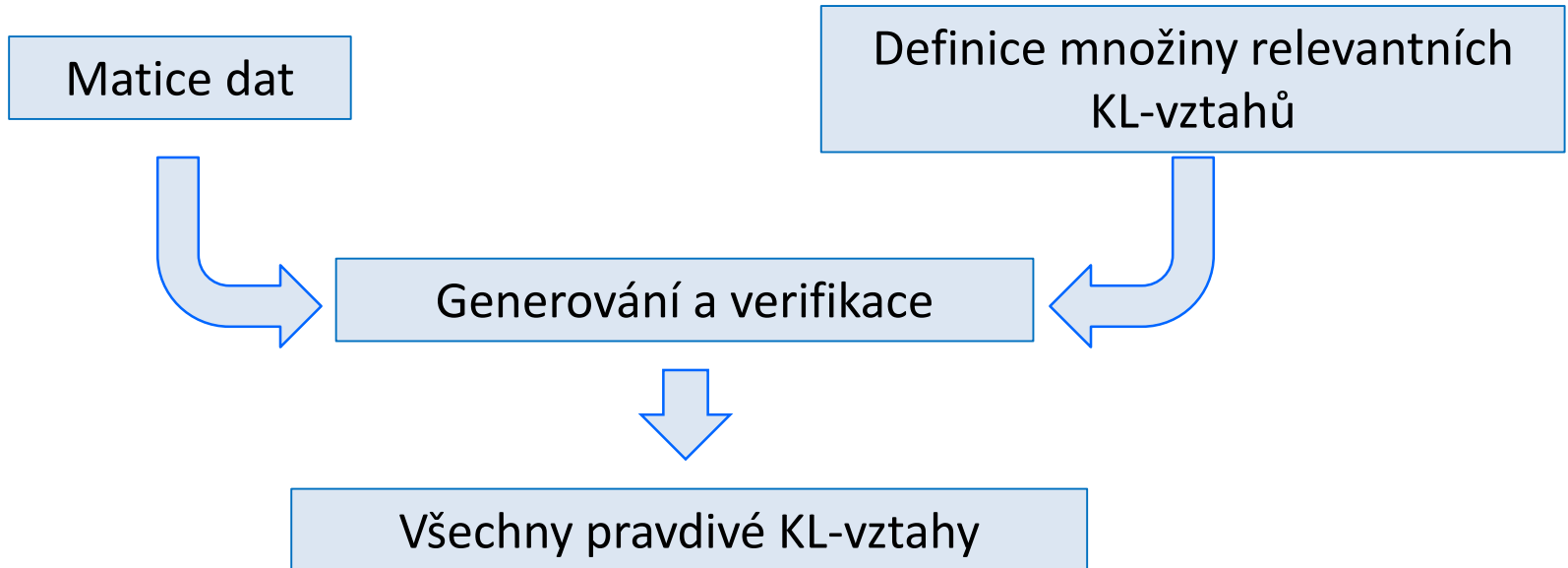
Týká se matice dat  $\mathcal{M}$



Je pravdivý: podmínka daná KL-kvantifikátorem je splněna na  $KL(R, C, \mathcal{M}/\chi)$

Je nepravdivý: podmínka daná KL-kvantifikátorem není plněna na  $KL(R, C, \mathcal{M}/\chi)$

# GUHA procedura KL-Miner



# Skupina atributů Dotazník

Matrix: HotelPlusExterni

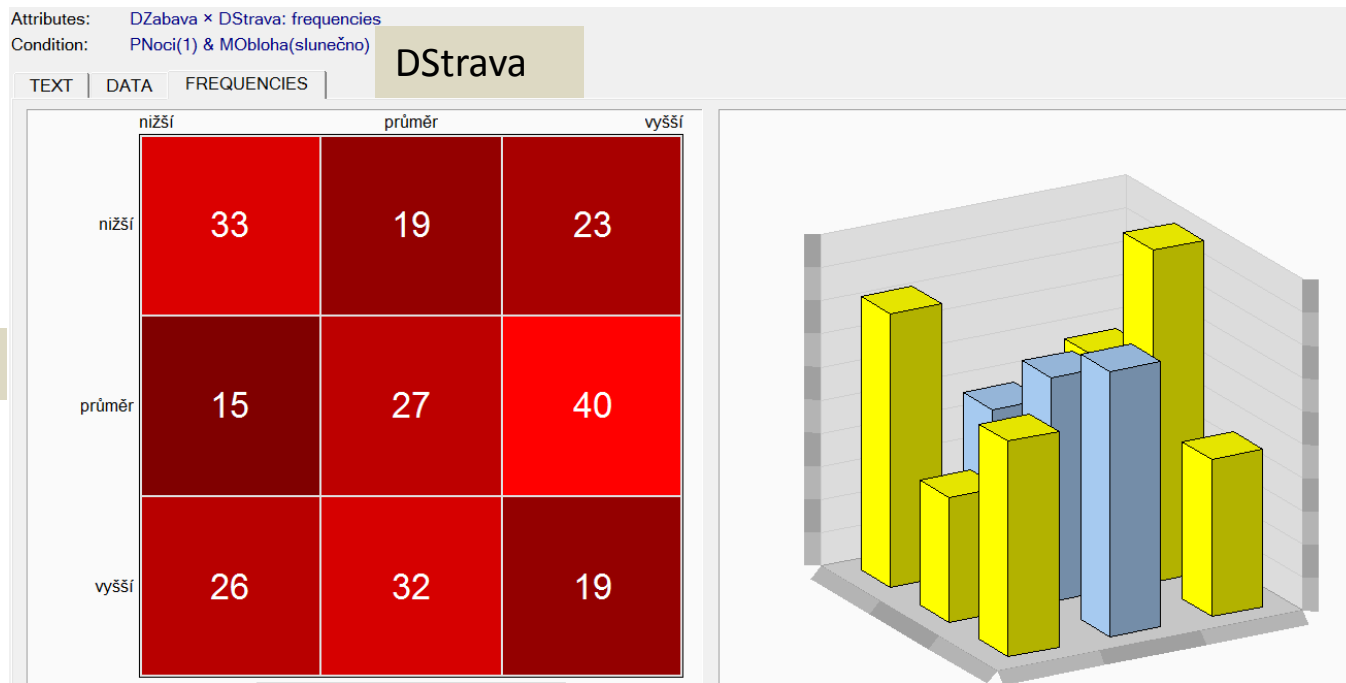
Groups of attributes tree

- Root group of attributes
  - Dotazník
  - Host
    - Bydliště
    - Meteo
  - Pobyt
    - Cena
    - Začátek
    - Směnárna

| Atribute          | Used | DBCColumn  | Categories | XCat | Sample categories                                       |
|-------------------|------|------------|------------|------|---|
| DHodnoceni ←      | +    | DHodnoceni | 3          |      | nespokojen, průměr, spokojen                            |
| DPersonal_edc20   | +    | DPersonal  | 20         |      | <0;4>, <5;9>, <10;14>, <15;19>, <20;24>, <25;29>, <30;3 |
| DPersonal_edc5    | +    | DPersonal  | 5          |      | <0;19>, <20;39>, <40;59>, <60;79>, <80;100>             |
| DPersonal_edc5_m  | +    | DPersonal  | 5          |      | , *, **, ***, ****, *****                               |
| DPersonal_ef3 ←   | +    | DPersonal  | 3          |      | nižší, průměr, vyšší                                    |
| DStrava_edc20     | +    | DStrava    | 20         |      | <0;4>, <5;9>, <10;14>, <15;19>, <20;24>, <25;29>, <30;3 |
| DStrava_edc5      | +    | DStrava    | 5          |      | <0;19>, <20;39>, <40;59>, <60;79>, <80;100>             |
| DStrava_edc5_m    | +    | DStrava    | 5          |      | , *, **, ***, ****, *****                               |
| DStrava_ef3 ←     | +    | DStrava    | 3          |      | nižší, průměr, vyšší                                    |
| DUbytovani_edc20  | +    | DUbytovani | 20         |      | <0;4>, <5;9>, <10;14>, <15;19>, <20;24>, <25;29>, <30;3 |
| DUbytovani_edc5   | +    | DUbytovani | 5          |      | <0;19>, <20;39>, <40;59>, <60;79>, <80;100>             |
| DUbytovani_edc5_m | +    | DUbytovani | 5          |      | , *, **, ***, ****, *****                               |
| DUbytovani_ef3 ←  | +    | DUbytovani | 3          |      | nižší, průměr, vyšší                                    |
| DZabava_edc20     | +    | DZabava    | 20         |      | <0;4>, <5;9>, <10;14>, <15;19>, <20;24>, <25;29>, <30;3 |
| DZabava_edc5      | +    | DZabava    | 5          |      | <0;19>, <20;39>, <40;59>, <60;79>, <80;100>             |
| DZabava_edc5_m    | +    | DZabava    | 5          |      | , *, **, ***, ****, *****                               |
| DZabava_ef3 ←     | +    | DZabava    | 3          |      | nižší, průměr, vyšší                                    |

# Aplikace procedury KL-Miner – příklad

Inspirace: KL-tabulka KL(DZabava, Dstrava, HotelPlusExterni / PNoci(1)  $\wedge$  Mobloha(oblačno))



Existuje kombinace hodnot atributů z Pobyť, Host, Bydliště, Meteo tak, že pro tuto kombinaci je zřetelná pozitivní ordinální závislost mezi některou dvojicí atributů z DHodnoceni, DPersonal, DStrava, DUbytování a DZabava?



# Závislost mezi atributy – příklad

KL(DZabava, DStrava, HotelPlusExterni /  
/ PNoci(1)  $\wedge$  MObloha(oblačno))

|        | nižší | průměr | vyšší |
|--------|-------|--------|-------|
| nižší  | 33    | 19     | 23    |
| průměr | 15    | 27     | 40    |
| vyšší  | 26    | 32     | 19    |

téměř nezávislé

KL(DHodnoceni, DPersonal, HotelPlusExterni /  
/ PNoci(3;7)  $\wedge$  PDenTydne(So))  $\wedge$  POsob(1,2)

|            | nižší | průměr | vyšší |
|------------|-------|--------|-------|
| nespokojen | 59    | 0      | 0     |
| průměr     | 2     | 25     | 5     |
| spokojen   | 0     | 0      | 65    |

téměř funkce

# Kendallův koeficient $\tau_b$

$$TauB(T_{KL}) = \frac{2(P-Q)}{\sqrt{(n^2 - \sum_{k=1}^K n_{k,*}^2)(n^2 - \sum_{l=1}^L n_{*,l}^2)}}$$

$$P = \sum_{k=1}^K \sum_{l=1}^L n_{k,l} \sum_{i=k+1}^K \sum_{j=l+1}^L n_{i,j} \quad Q = \sum_{k=1}^K \sum_{l=1}^L n_{k,l} \sum_{i=k+1}^K \sum_{j=1}^{l-1} n_{i,j}$$

$$T_{KL} :$$

| $\mathcal{M}/\chi$ | $c_1$     | $\dots$ | $c_L$     | $\Sigma_l$ |
|--------------------|-----------|---------|-----------|------------|
| $r_1$              | $n_{1,1}$ | $\dots$ | $n_{1,L}$ | $n_{1,*}$  |
| $\vdots$           | $\vdots$  |         | $\vdots$  | $\vdots$   |
| $r_K$              | $n_{K,1}$ | $\dots$ | $n_{K,L}$ | $n_{K,*}$  |
| $\Sigma_k$         | $n_{*,1}$ | $\dots$ | $n_{*,L}$ | $n$        |

$KL(R, C, \mathcal{M}/\chi)$

$TauB(T_{KL}) > 0$  – pozitivní ordinální závislost pokud platí  $\chi$ ,  $R \uparrow \uparrow C / \chi$

$TauB(T_{KL}) < 0$  – negativní ordinální závislost pokud platí  $\chi$ ,  $R \uparrow \downarrow C / \chi$

$TauB(T_{KL}) = 0$  – ordinální nezávislost  $R$  a  $C$  pokud platí  $\chi$

$|TauB(T_{KL})| = 1 - C$  je funkcí  $R$  a  $C$  pokud platí  $\chi$

# Závislost mezi atributy – příklad s $\tau_b$

KL(DZabava, DStrava, HotelPlusExterni /  
/ PNoci(1)  $\wedge$  MObloha(oblačno)

|        | nižší | průměr | vyšší |
|--------|-------|--------|-------|
| nižší  | 33    | 19     | 23    |
| průměr | 15    | 27     | 40    |
| vyšší  | 26    | 32     | 19    |

téměř nezávislé

$$\tau_b = 0.01$$

KL(DHodnoceni, DPersonal, HotelPlusExterni /  
/ PNoci(3;7)  $\wedge$  PDenTydne(So))  $\wedge$  POsob(1,2)

|            | nižší | průměr | vyšší |
|------------|-------|--------|-------|
| nespokojen | 59    | 0      | 0     |
| průměr     | 2     | 25     | 5     |
| spokojen   | 0     | 0      | 65    |

téměř funkce

$$\tau_b = 0.96$$

# Aplikace procedury KL-Miner – příklad 1

| HVek | HPohlavi | HMesto           | HMesto_X   | HMesto_Y   | HStat     | PPobytOd   | PNoci | POsob | PTypPobytu | PCenaUbytovani | PCenaStrava | PCenaSleva | PCenaCelkem | DHodnoceni             |
|------|----------|------------------|------------|------------|-----------|------------|-------|-------|------------|----------------|-------------|------------|-------------|------------------------|
| 21   | žena     | České Budějovice | 14.4757883 | 48.9763169 | ČR        | 31.5.2013  | 1     | 1     | rekreační  | 1450.00        | 0.000       | 0.00       | 1450.00     | spokojen 91 82 81 56   |
| 34   | muž      | Linec            | 14.2862742 | 48.3066489 | Rakousko  | 2.8.2013   | 2     | 4     | rekreační  | 11600.00       | 1440.000    | 200.00     | 12840.00    | průměr 44 21 62 84     |
| 30   | muž      | Videň            | 16.3736767 | 48.2115631 | Rakousko  | 5.6.2012   | 7     | 2     | rekreační  | 16940.00       | 2100.000    | 200.00     | 18840.00    | nespokojen 5 37 25 71  |
| 62   | muž      |                  | 3          | 50.        |           |            |       |       |            |                |             |            |             |                        |
| 35   | žen      |                  | 48.21      |            |           |            |       |       |            |                |             |            |             |                        |
| 58   | muž      |                  | 4.4757     |            |           |            |       |       |            |                |             |            |             |                        |
| 81   | žena     | Videň            | 16.3736767 | 48.21      |           |            |       |       |            |                |             |            |             |                        |
| 22   | žena     | Dráždany         | 13.7397044 | 5          |           |            |       |       |            |                |             |            |             |                        |
| 82   | muž      | Katovice         | 19.0241283 | 50.        |           |            |       |       |            |                |             |            |             |                        |
| 55   | muž      | Praha            | 14.4212806 | 50.0874    |           |            |       |       |            |                |             |            |             |                        |
| 75   | žena     | Berlín           | 13.3908886 | 52.        |           |            |       |       |            |                |             |            |             |                        |
| 66   | žena     | Linec            | 14.2862742 | 48.30      |           |            |       |       |            |                |             |            |             |                        |
| 64   | žena     | Linec            | 14.2862742 | 48.30      |           |            |       |       |            |                |             |            |             |                        |
| 35   | muž      | Košice           | 21.2543528 | 48.7160408 | Slovensko | 6.4.2012   | 1     | 2     | rekreační  | 2420.00        | 0.000       | 0.00       | 2420.00     | průměr 36 52 56 36     |
| 32   | muž      | Mnichov          | 11.5836375 | 48.1364669 | Německo   | 13.8.2013  | 1     | 1     | služební   | 1450.00        | 180.000     | 0.00       | 1630.00     | průměr 47 58 70 46     |
| 65   | muž      | Plzeň            | 13.3771556 | 49.7490    |           |            |       |       |            |                |             |            |             |                        |
| 79   | muž      | Brno             | 16.6153758 | 49.1921    |           |            |       |       |            |                |             |            |             |                        |
| 28   | žena     | Dráždany         | 13.7397044 | 5          |           |            |       |       |            |                |             |            |             |                        |
| 35   | žena     | Hamburg          | 10.0043528 | 53.5498325 | Německo   | 5.1.2013   | 14    | 2     | rekreační  | 40600.00       | 5040.000    | 200.00     | 45440.00    | nespokojen 27 15 30 12 |
| 22   | muž      | Plzeň            | 13.3771556 | 49.7490406 | ČR        | 9.11.2013  | 4     | 1     | rekreační  | 5800.00        | 0.000       | 200.00     | 5600.00     | spokojen 84 79 86 48   |
| 25   | muž      | Karlovy Vary     | 12.8690381 | 50.2311075 | ČR        | 9.11.2013  | 7     | 2     | rekreační  | 20300.00       | 0.000       | 600.00     | 19700.00    | spokojen 88 74 94 54   |
| 20   | žena     | Hamburg          | 10.0043528 | 53.5498325 | Německo   | 19.1.2013  | 14    | 4     | rekreační  | 81200.00       | 10080.000   | 200.00     | 91080.00    | průměr 59 23 46 96     |
| 30   | žena     | Linec            | 14.        |            |           | .11.2013   | 2     | 4     | rekreační  | 11600.00       | 0.000       | 200.00     | 11400.00    | spokojen 88 82 84 61   |
| 45   | žena     | Karlovy V        |            |            |           | 28.12.2012 | 2     | 2     | rekreační  | 4840.00        | 600.000     | 200.00     | 5240.00     | spokojen 81 99 87 23   |

Hotel.txt

Existuje kombinace hodnot atributů z Pobyt, Host, Bydliště, Meteo tak, že pro tuto kombinaci existuje pozitivní ordinální závislost mezi některou dvojicí atributů ze skupiny Dotazník ( DHodnoceni, DPersonal, DStrava, DUbytování a Dzabava)?

? KL:  $\uparrow\uparrow_{0.9}$  Dotazník / Pobyt, Host, Bydliště, Meteo

Meteo.txt

| MDatum    | MTeplota | MOblaha    |
|-----------|----------|------------|
| 4.1.2012  | -6.3     | slunečno   |
| 5.1.2012  | -6.6     | zataženo   |
| 6.1.2012  | 6.1      | srážky     |
| 7.1.2012  | 1.6      | srážky     |
| 8.1.2012  | -1.3     | srážky     |
|           |          | 7 zataženo |
|           |          | srážky     |
|           |          | zataženo   |
| 12.1.2012 | -3.1     | srážky     |
| 13.1.2012 | -8.1     | zataženo   |
| 14.1.2012 | -10.7    | srážky     |
| 15.1.2012 | -5.5     | zataženo   |
| 16.1.2012 | 2.3      | zataženo   |
| 17.1.2012 | -1.9     | zataženo   |
| 18.1.2012 | -8.6     | zataženo   |

# KL: $\uparrow\uparrow_{0.9}$ Dotazník / Pobyt, Host, Bydliště, Meteo

The screenshot shows the 'Data-mining Task basic parameters' window for a task named '02: Dotaznik\_ef3 x Dotaznik\_ef3 / Pobyt, Host, Bydliště, Meteo'. The task type is 'KL-Miner' and the data matrix is 'HotelPlusExterni'. The window is divided into several panes:

- ROW ATTRIBUTES:** Lists row attributes such as DHodnoceni, DPersonal\_ef3, DStrava\_ef3, DUbytovani\_ef3, and DZabava\_ef3. A green box highlights the word 'Dotazník'.
- QUANTIFIERS:** A table showing quantifiers used in the task:
 

| Type | Range | Rel. | Value  | Units |
|------|-------|------|--------|-------|
| SUM  | all   | >=   | 250.00 | Abs   |
| KEND | all   | >=   | 0.90   | Abs   |

 A green box highlights the expression  $SUM \geq 250 \wedge \tau_b \geq 0.9$ .
- COLUMN ATTRIBUTES:** Lists column attributes similar to the row attributes. A green box highlights the word 'Dotazník'.
- CONDITION:** Lists generated conditions for various attributes like Pobyt, Host, Bydliště, and Meteo. A green box highlights the expression  $Pobyt (*) \wedge Host (*) \wedge Bydliště(*) \wedge Meteo(*)$ .
- Aggregate function:** Shows 'Type: Count(\*)' and 'Attribute: -'.
- Task parameters:** Includes options for 'Include 'worse' extensions of condition' (Yes), 'Extensions minimal length check' (Yes), and 'Include extensions of coefficients with no change in the histogram' (Yes).

A blue box at the bottom contains the text: 'Jedno z více možných zadání pro řešení dané analytické otázky'.

# KL-kvantifikátor SUM

| QUANTIFIERS |       |      |        |       |
|-------------|-------|------|--------|-------|
| Type        | Range | Rel. | Value  | Units |
| SUM         | all   | >=   | 250.00 | Abs   |
| KEND        | all   | >=   | 0.90   | Abs   |

KL Simple frequencies quantifier settings

Interest measure type: **Sum of frequencies**

Sum of frequencies from given part of contingency table

Relation: Greater than or equal

Threshold value: 250

Threshold-value units: Absolute number

Category Range:

|         |      |     |
|---------|------|-----|
|         | From | To  |
| Rows    | 0    | 100 |
| Columns | 0    | 100 |

Absolute category index  
 Relative range [%] to act number of categories

Primary IM Settings:

|   |      |    |
|---|------|----|
| <input type="checkbox"/> Set as primary IM      | From | To |
| <input type="checkbox"/> Normalize value range: | 0    | 1  |

Note: -

OK Cancel

SUM >= 250 minimálně 250 řádků splňuje podmínku

# Kendallův kvantifikátor

| QUANTIFIERS |       |      |        |       |
|-------------|-------|------|--------|-------|
| Type        | Range | Rel. | Value  | Units |
| SUM         | all   | >=   | 250.00 | Abs   |
| KEND        | all   | >=   | 0.90   | Abs   |

KL Statistical quantifier settings

Interest measure type: Kendall's TauB coefficient  
Kendall's TauB (coefficient of ordinal correlation of two variables) in  $<-1;1>$  (the farther is value from 0 the more dependant)

Relation: Greater than or equal

Threshold value: 0.9

Category Range: From 0 To 100  
Columns: 0

Absolute category  
 Relative range [%] to act number of categories

Parameters:  Absolute value of TauB for Kendall's coefficient (ie. interval  $<0;1>$  only)

Formula:

Note:

OK Cancel

**Kendallův koeficient  $\tau_b \geq 0.9$**

# KL-Miner – příklad výstupu

Task run

Start: 27.2.2016 20:49:33      Total time: 0h 0m 1s

Number of verifications: 45760      Mode: Standard

Number of hypotheses: 54

Add group    Del group    Edit group

Actual group of hypotheses: All hypotheses

Hypotheses in group: 54      Shown hypotheses: 54      Highlighted: 0

| Nr. | Id | TauB  | Hypothesis   |
|-----|----|-------|--|
| 1   | 37 | 0.937 | DHodnoceni × DPersonal / PDenTydne(So) & POsob(<=2) & HVek(>=od 28 do 60)                          |
| 2   | 38 | 0.937 | DPersonal × DHodnoceni / PDenTydne(So) & POsob(<=2) & HVek(>=od 28 do 60)                          |
| 3   | 35 | 0.934 | DHodnoceni × DPersonal / PDenTydne(So) & POsob(<=2)  |
| 4   | 36 | 0.934 | DPersonal × DHodnoceni / PDenTydne(So) & POsob(<=2)  |
| 5   | 23 | 0.919 | DHodnoceni × DPersonal / PDenTydne(So) & MTeplota(zima,neutrální)                                  |
| 6   | 24 | 0.919 | DPersonal × DHodnoceni / PDenTydne(So) & MTeplota(zima,neutrální)                                  |
| 7   | 1  | 0.919 | DHodnoceni × DPersonal / PNoci(<3;6>,7) & PDenTydne(So)  |
| 8   | 2  | 0.919 | DPersonal × DHodnoceni / PNoci(<3;6>,7) & PDenTydne(So)  |
| 9   | 33 | 0.919 | DHodnoceni × DPersonal / PDenTydne(So) & HVek(>=od 28 do 60) & MTeplota(zima,neutrální)            |
| 10  | 34 | 0.919 | DPersonal × DHodnoceni / PDenTydne(So) & HVek(>=od 28 do 60) & MTeplota(zima,neutrální)            |
| 11  | 11 | 0.917 | DHodnoceni × DPersonal / PNoci(7,<8;13>) & PDenTydne(So)   |
| 12  | 12 | 0.917 | DPersonal × DHodnoceni / PNoci(7,<8;13>) & PDenTydne(So)   |
| 13  | 51 | 0.916 | DHodnoceni × DPersonal / PDenTydne(So) & OSobonoci(>=vyšší) & MTeplota(zima,neutrální)             |
| 14  | 52 | 0.916 | DPersonal × DHodnoceni / PDenTydne(So) & OSobonoci(>=vyšší) & MTeplota(zima,neutrální)             |
| 15  | 49 | 0.916 | DHodnoceni × DPersonal / PDenTydne(So) & OSob(>=2) & OSobonoci(>=vyšší) & MTeplota(zima,neutrální) |
| 16  | 50 | 0.916 | DPersonal × DHodnoceni / PDenTydne(So) & OSob(>=2) & OSobonoci(>=vyšší) & MTeplota(zima,neutrální) |
| 17  | 9  | 0.916 | DHodnoceni × DPersonal / PNoci(<3;6>,7) & OSobonoci(průměr,vyšší)                                  |
| 18  | 10 | 0.916 | DPersonal × DHodnoceni / PNoci(<3;6>,7) & OSobonoci(průměr,vyšší)                                  |
| 19  | 7  | 0.915 | DHodnoceni × DPersonal / PNoci(<3;6>,7) & OSob(<=3) & OSobonoci(průměr,vyšší)                      |

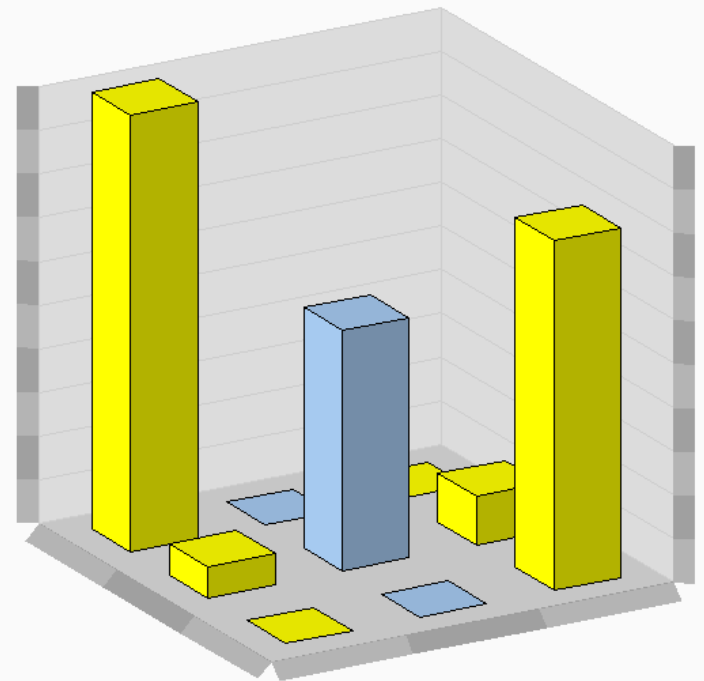
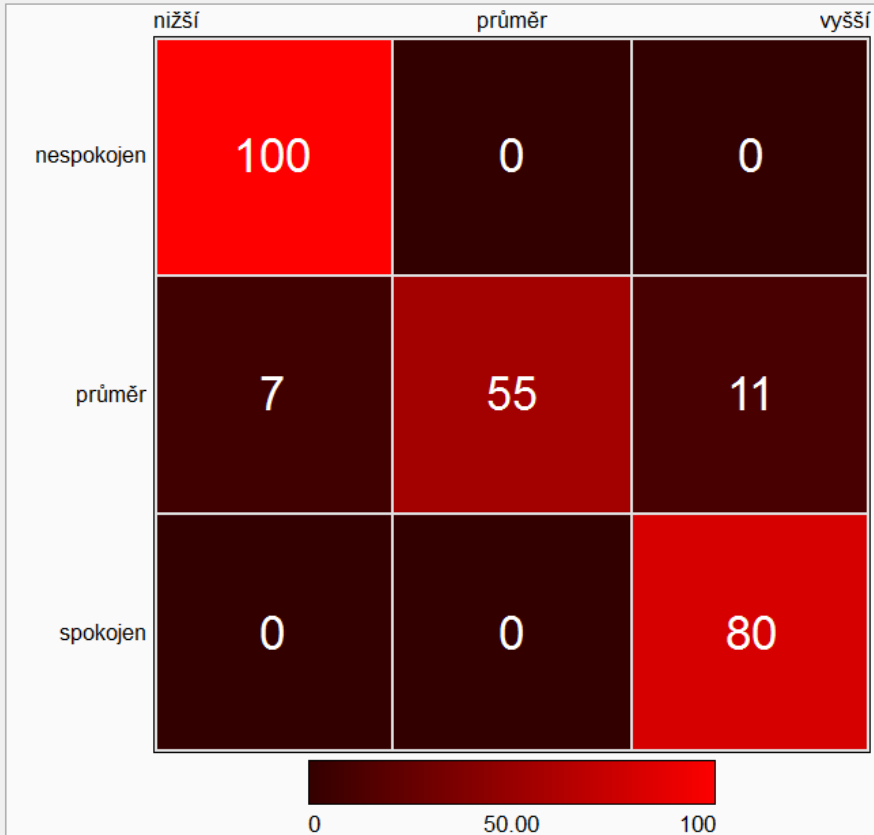


# KL-Miner – příklad detailního výstupu (1)

DHodnoceni  $\uparrow\uparrow$  DPersonal / PDenTydne(So)  $\wedge$  Posob(1,2)  $\wedge$  HVek(28 a vyšší)  
SUM = 253,  $\tau_b = 0.94$

Attributes: DHodnoceni  $\times$  DPersonal: frequencies  
Condition: PDenTydne(So) & POsob(1, 2) & HVek(od 28 do 60, 60 a více)

TEXT | DATA | FREQUENCIES



# KL-Miner – příklad detailního výstupu (2)

DHodnoceni  $\uparrow\uparrow$  DPersonal / PDenTydne(So)  $\wedge$  HVek(28 a vyšší)  
SUM = 342,  $\tau_b = 0.90$

Attributes: DHodnoceni  $\times$  DPersonal: frequencies  
Condition: PDenTydne(So) & HVek(od 21 do 28, od 28 do 60)

TEXT | DATA | FREQUENCIES

|            | nižší | průměr | vyšší |
|------------|-------|--------|-------|
| nespokojen | 122   | 0      | 0     |
| průměr     | 16    | 99     | 18    |
| spokojen   | 0     | 2      | 85    |

