35. Porozumění datům (důvod a cíl; popisné charakteristiky dat a vizualizační techniky; korelační analýza).

Porozumění datům je typicky druhou fází analytického projektu dle modelu CRISP-DM, nastává po porozumění kontextu. Hlavním cílem je dozvědět se co nejvíce o dostupných datech. Předpokladem pro začátek této fáze je, že chápeme zkoumanou problematiku, máme stanovený cíl, resp. konkrétní datamining úlohu. Vstupem této fáze jsou datové zdroje a dokumentace k nim, výstupem pak informace o vlastnostech dat (popisné charakteristiky, grafy) a podklady pro vytvoření datové sady, která následně bude využita pro analýzu dat. Fáze porozumění zahrnuje:

1. Rozpracování informace k dostupným datům – jaká data máme, jejich věrohodnost
2. Popis dat – struktura, formát (pokud už není součástí dodané dokumentace)
3. Prozkoumání dat (explorační analýza) – popisné charakteristiky, grafy, korelace
4. Zhodnocení kvality dat – chybí nějaké hodnoty? jsou data zašuměná?

Datové sady

Datová sada je kolekce dat vybraných k dolování. Datová sada se typicky skládá z datových objektů a jednotlivé objekty jsou popsány atributy (každý atribut popisuje nějakou vlastnost/rys objektu). Dataset tradičních dat (např. CSV, viz otázka 34 – zdroje a typy dat) má obvykle podobu tabulky (flat file), ve které řádky obsahují jednotlivé datové objekty a sloupce pak reprezentují atributy. Obvykle jsou datovými objekty entity reálného světa, atributy mají tedy pochopitelnou sémantiku (např. věk osoby), nemusí to tak ale být nutně vždy (např. vektor reprezentující obrázek).

Rozlišujeme několik typů atributů:

- Kategorický – řetězec nebo celočíselná hodnota (kvalitativní)
  - Binární – nominální se dvěma stavy
    - Symetrický binární – oba stavy stejně důležité, např. pohlaví Male/Female
    - Asymetrický binární – stavy nemají stejnou důležitost, např. test Pozitivní/Negativní
  - Nominální – kategorie, stavy, jména věcí. Např. barva z {černá, modrá, bílá}
  - Ordinální – je definováno uspořádání nad možnými hodnotami atributu, ale rozdíl dvou sousedních hodnot není definován, např. vzdělání z {základní, střední všeobecné, střední odborné, vysokoškolské}
- Numerický/kvantitativní – celočíselný nebo reálný
  - Intervalový

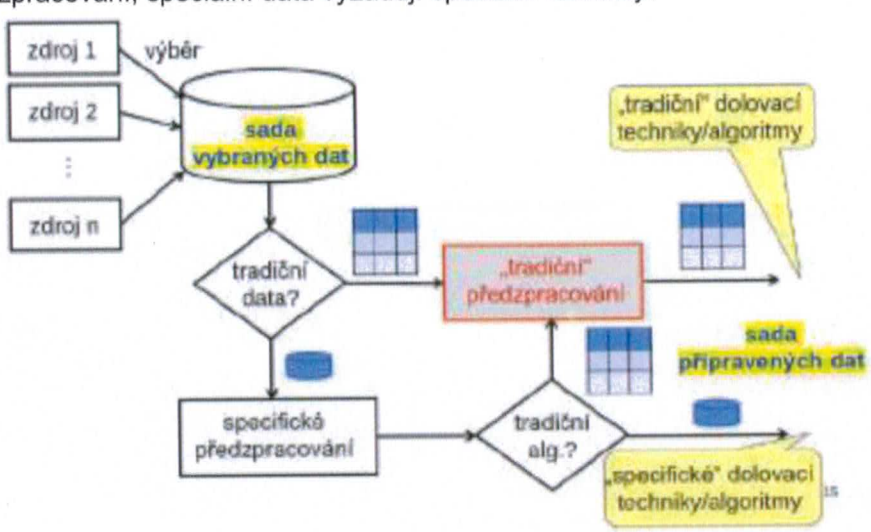
-10°C / 5°C ← nic to neříká, nemá nulový bod, intervalový numerický atribut

- hodnoty jsou uspořádatelné, jsme schopni určit rozdíl, ale ne podíl
- chybí nulový bod
- např. teplota ve stupních Celsia, datum

○ Poměrový

- má implicitní nulový bod, jen kladné hodnoty
- jsme schopni určit i podíl dvou hodnot
- teplota v Kelvinech, počet dětí, počet peněz

Práce s datovou sadou se řídí následujícím schématem – tradiční data vedou na tradiční zpracování, speciální data vyžadují speciální techniky:



Charakteristiky dat

U datové sady chceme zjistit typicky rozložení hodnot atributů. K tomu slouží popisné charakteristiky dat. Rozlišujeme několik druhů charakteristik:

- míry polohy – určují střed dat
- míry variability – určuji rozptýlenost hodnot kolem středu
- další, např. šikmost, špičatost

Dále rozlišujeme charakteristiky podle možnosti distribuce výpočtu:

- distributivní – výpočet lze zcela distribuovat, např. počet prvků
- algebraická – výsledek je algebraickou operací nad jednou nebo více distributivními mírami (např. průměr – počet a součet lze určit distributivně)
- holistická – nelze distributovat, je třeba udělat výpočet nad celým souborem, např. medián

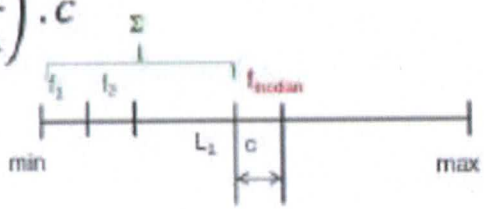
Míry polohy

- **Aritmetický průměr** (střední hodnota): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\mu = \frac{\sum x}{N}$
  - ***Vážený průměr:*** $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - citlivost na odlehlé hodnoty
  - ***Upravený průměr (trimmed mean)***: ořezání extrémních hodnot
- **Geometrický průměr** - pro *n* kladných hodnot:

  $$\overline{x_G} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

  - Používá se typicky pro výpočet průměrného tempa růstu
- **Harmonický průměr** - pro *n* kladných hodnot: $\overline{x_H} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$
  - Pro hodnoty výrazně nesymetricky rozptýlené kolem středu
  - Převrácená hodnoty průměru převrácených hodnot
- **Medián** (holistická míra):
  - Prostřední hodnota seřazeného seznamu, je-li počet hodnot lichý, jinak průměr dvou prostředních hodnot
  - Odhad interpolací (pro *n* hodnot seskupených do intervalů):

    $$median = L_1 + \left(\frac{\frac{n}{2} - \sum_l f_l}{f_{median}}\right).c$$

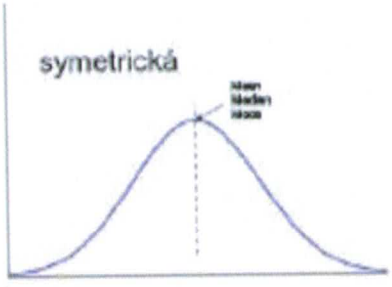


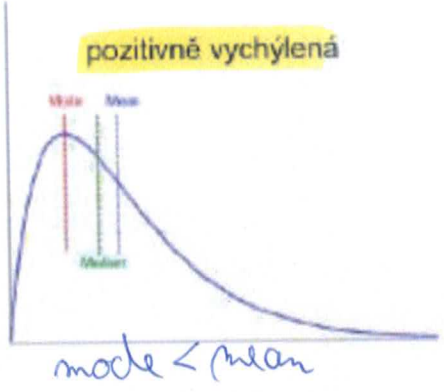
- **Modus:**
  - Nejčastější hodnota v datech
  - Jedno-, dvou-, ..., (multimodální)
  - Empirická formule pro málo vychýlená data:

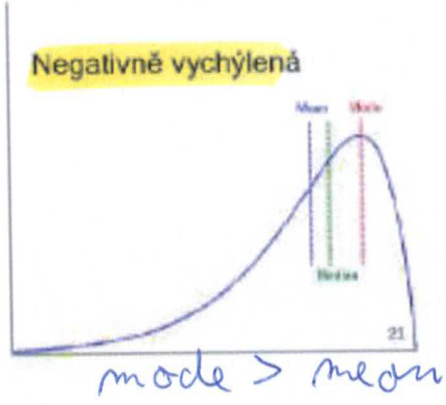
    $$\bar{x} - modus(x) = 3(\bar{x} - median(x))$$

Symetrická a vychýlená data

- Medián, modus a průměr u symetrických dat splývají







Kvantil

k-tý q-kvantil je hodnota q_k náhodné veličiny X, pro kterou platí $P(X \leq q_k) \geq k/q \wedge P(X \geq q_k) \geq 1 - k/q$. (q-1) q-kvantilů rozděluje uspořádaný soubor na q přibližně stejně početných intervalů. Například pro q=4 hovoříme o tzv. kvartilech, máme první, druhý a třetí kvartil. První kvartil (q_1) je hodnota, která je větší než čtvrtina hodnot a menší než tři čtvrtiny hodnot. Druhý kvartil je ekvivalentní mediánu. Kromě kvartilů se využívají i percentily (q = 100), decily (q=10) nebo kvintily (q=5).

Míry variability

- **Rozpětí** (variační šíře): $R = x_{max} - x_{min}$
  - ovlivněná extrémními hodnotami
- **Mezikvartilové rozpětí** (IQR - Interquartile range)
  - $IQR = q_3 - q_1$
  - 50% *středních* hodnot, **odlehlá hodnota** x:
    $(x < q_1 - 1.5IQR) \cup (x > q_3 + 1.5IQR)$
- **Rozptyl** (algebraická míra)

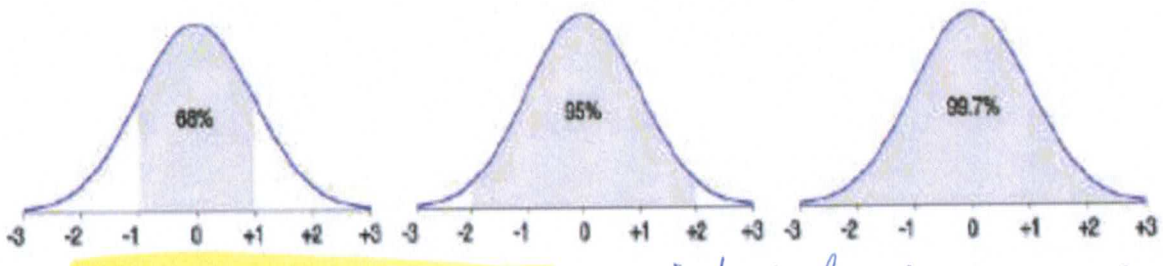
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \text{vzorek}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i{}^2 - \overline{x}^2 \quad \text{populace}$$

- **Směrodatná odchylka**

$$s = \sqrt{s^2} \qquad \sigma = \sqrt{\sigma^2}$$

- Rozptyl normálního rozdělení dat
  - pro křivku hustoty normálního rozdělení platí:
    - V intervalu <μ−σ, μ+σ> leží asi 68% hodnot
    - V intervalu <μ−2σ, μ+2σ> leží asi 95% hodnot
    - V intervalu <μ−3σ, μ+3σ> leží asi 99.7% hodnot



- pravidlo 3 sigma - v intervalu $(\mu - 3\sigma, \mu + 3\sigma)$ leží 99,7% hodnot

5/2

- **Průměrná absolutní odchylka**
  (*AAD – Average Absolute Deviation*)
  - *střední* hodnota absolutních hodnot odchylky od *středu*
  - obecně může být vztažena i k jinému *středu*
  - **Mediánová absolutní odchylka** (*median absolute deviation*)
    $MAD = median(|x_i - median|)$
    - robustní míra málo ovlivněná odlehlými hodnotami
  - **(Průměrná) absolutní odchylka** (*mean absolute deviation*)
    $MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$
- **Variační koeficient**
  $k = \frac{s}{\overline{x}}$
- **Sumarizace 5 čísly** - složená míra
  $< min, q_1, median, q_3, max >$

Vizualizační techniky

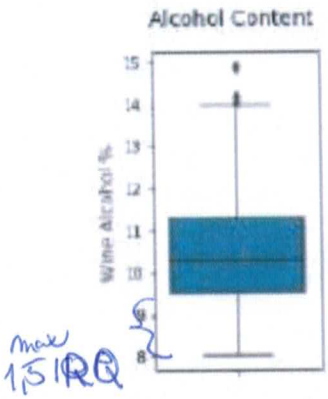
Grafy pro jeden atribut

Cílem těchto grafů je ukázat rozložení hodnot daného atributu. Pro kvantitativní atributy využíváme krabicové grafy, histogramy s rozdělením do košů, graf hustoty, houslový graf nebo kvantilový graf. Pro kategorické atributy se většina těchto grafů nehodí a vhodný je např. histogram.

6/2

Krabicový graf (boxplot)

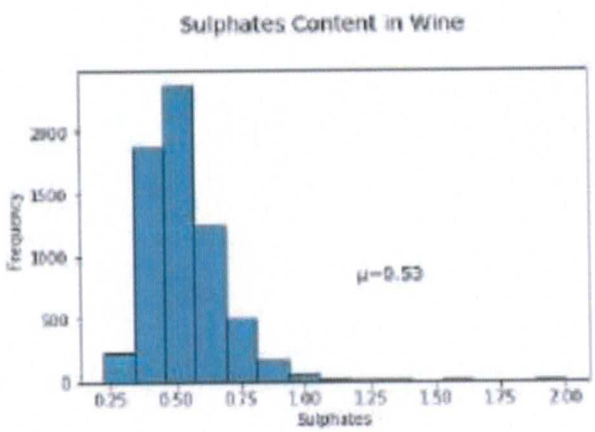
- Vizualizuje sumarizaci 5 čísly
- Data reprezentována obdélníkem
- Konce obdélníku jsou q_1 a q_3, tj. výška je IRQ
- Medián je vyznačen úsečkou uvnitř
- Protažení: dvě úsečky k minimu a maximu, případně individuální odlehlé hodnoty

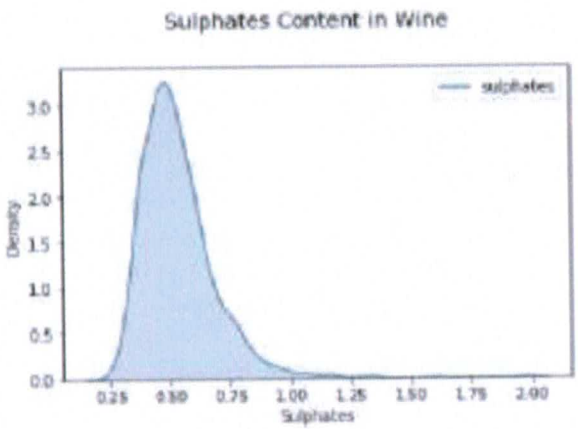


max 1,5 IRQ

Histogram a graf hustoty

- Histogram s rozdělením na intervaly (bins)
- Graf hustoty (density/kernel density) – zobrazuje odhad funkce hustoty pravděpodobnosti metodou KDE (Kernel Density Estimation).

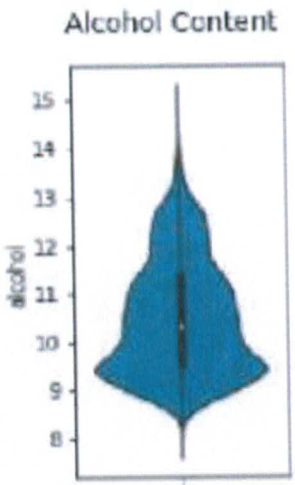


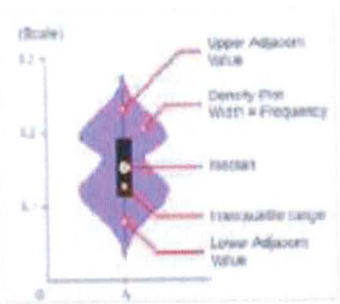


7/2

Houslový graf (violin plot)

- Podobný krabicovému, ale ukazuje navíc odhad hustoty pravděpodobnosti.
  - Vhodný zejména pro multimodální data.

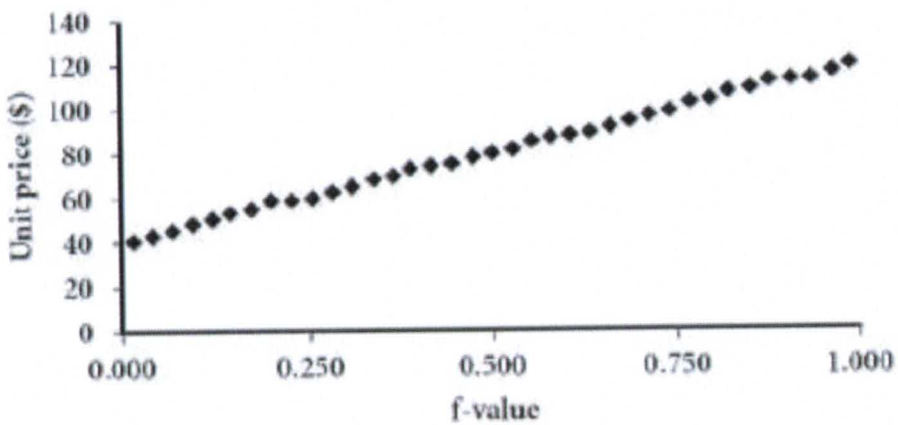




Zdroj [5]

Kvantilový graf

- Zobrazuje všechna data (možnost posoudit celkové chování i neobvyklé výskyty)
- Vykresluje informaci o kvantilech
  - f-hodnota mapuje index hodnoty x_i souboru dat uspořádaného vzestupně na hodnotu pravděpodobnosti, tj. hodnota f_i udává, že přibližně $100\,f_i$ % dat má hodnotu menší nebo rovnou x_i

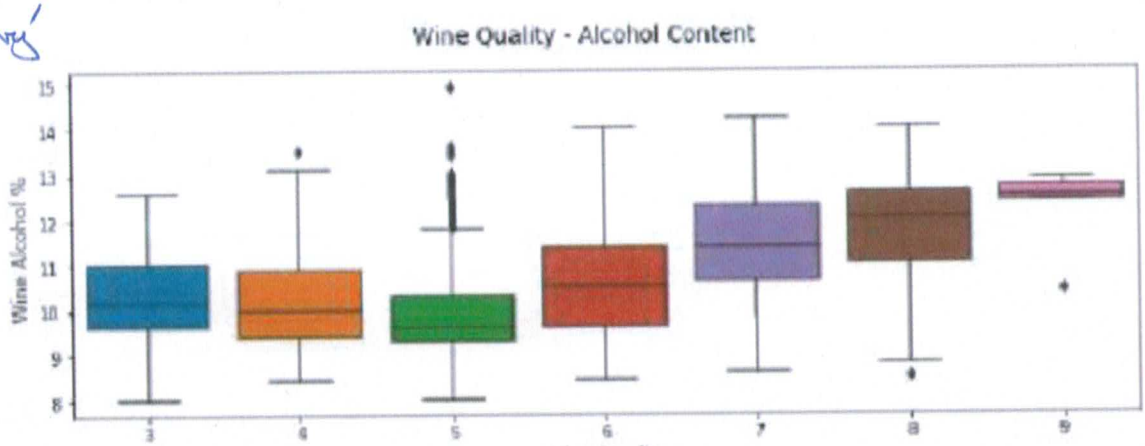


Grafy pro více atributů

Cílem vizualizace několika atributů typicky bývá porovnat rozložení hodnot nebo odhalit potenciální vztahy mezi atributy (korelace). Některé typy grafů pro jeden atribut lze rozšířit pro vizualizaci více atributů, typicky např. krabicový nebo houslový graf umožňuje vizualizovat zaráz jeden kvantitativní a jeden kategorický atribut:

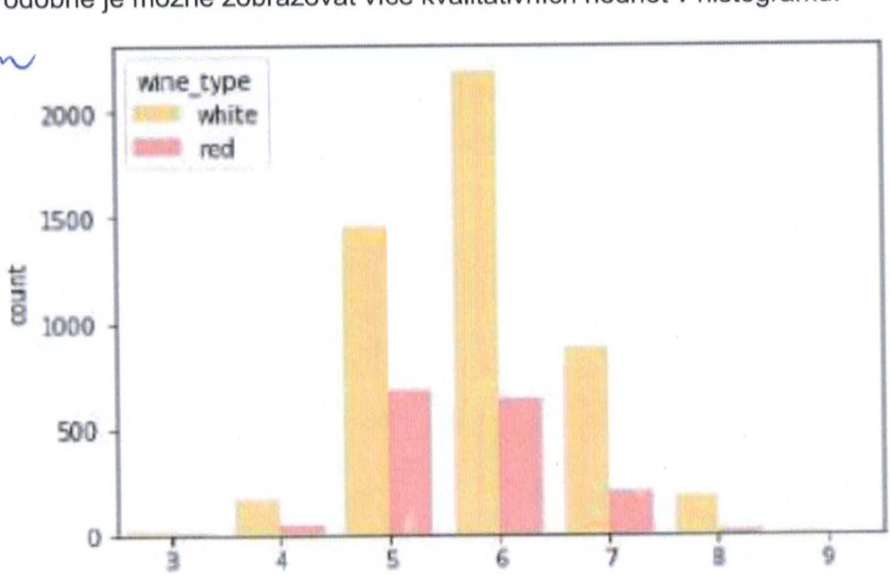
Krabicový

Pozn: jde stejně tak udělat i houslový



Podobně je možné zobrazovat více kvalitativních hodnot v histogramu:

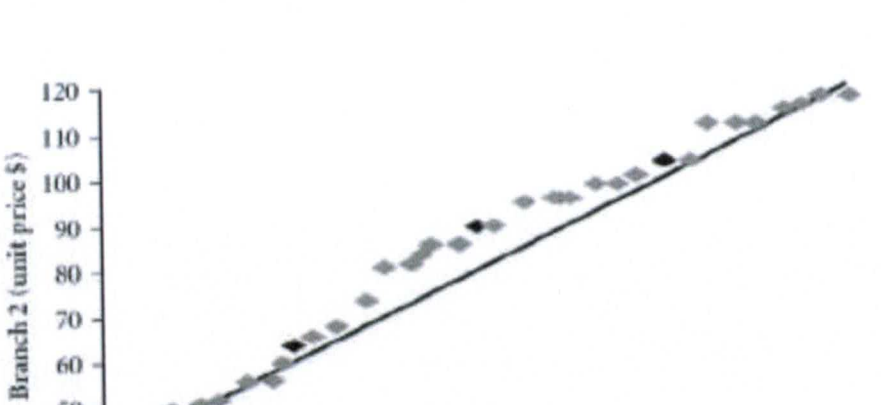
Histogram



9/2

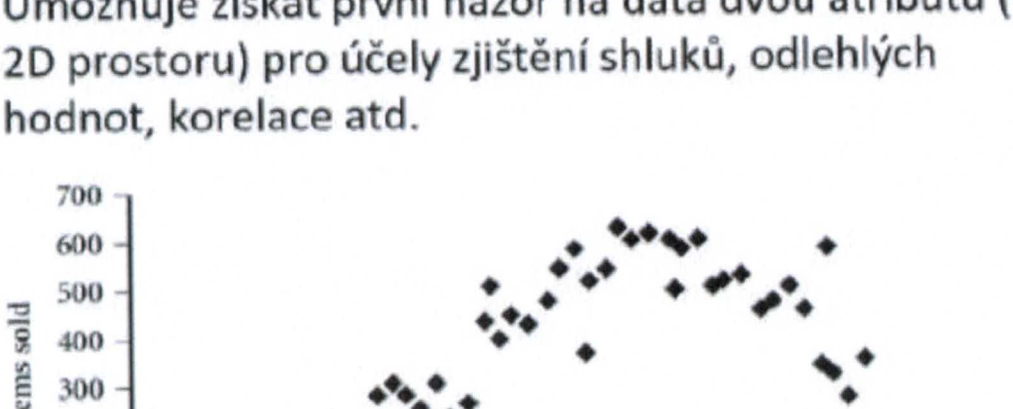
Q-Q graf

- Zobrazuje kvantily dvou atributů (proměnných) navzájem, tj bod grafu odpovídá stejné pravděpodobnosti na křivce distribuční funkce každého z atributů
- Umožňuje sledovat vzájemný posun distribuce hodnot obou atributů



Bodový graf

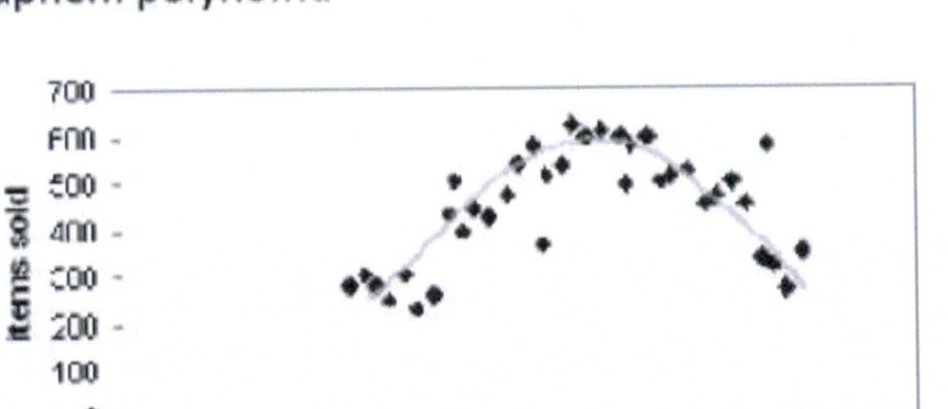
- Umožňuje získat první názor na data dvou atributů (v 2D prostoru) pro účely zjištění shluků, odlehlých hodnot, korelace atd.



10/2

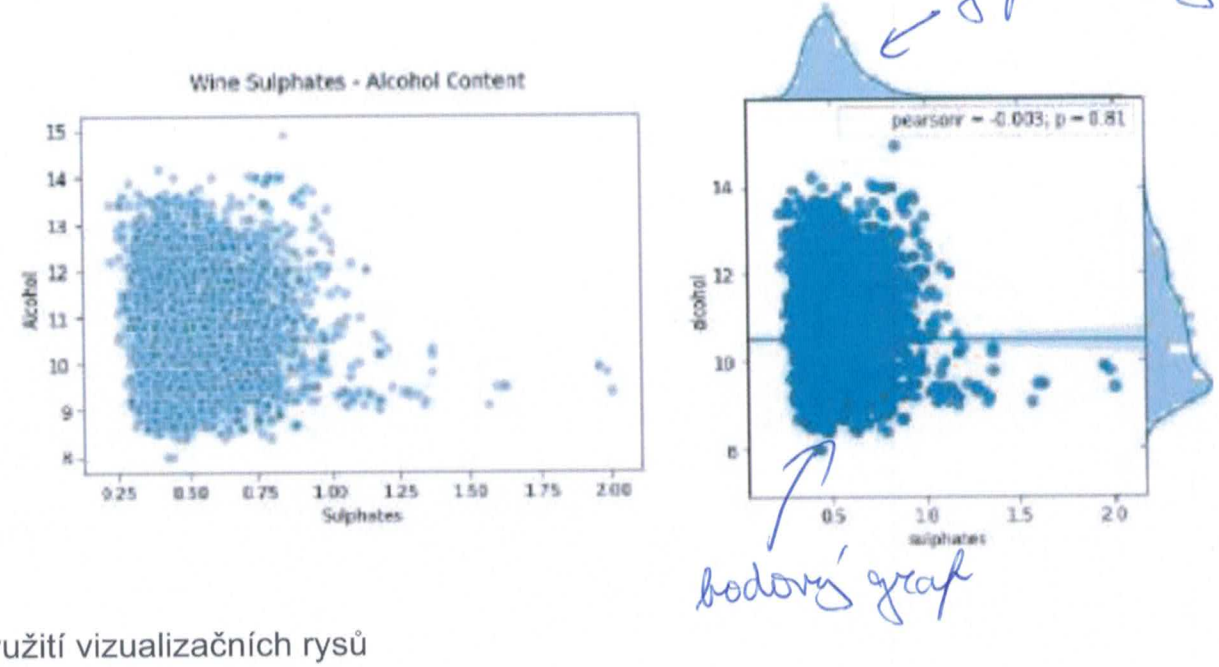
Křivka lokální regrese

- Přidá do bodového grafu vyhlazující křivku pro lepší zjištění závislosti atributů
- Vyhlazení je řízeno dvěma parametry: parametrem vyhlazení a stupněm polynomu



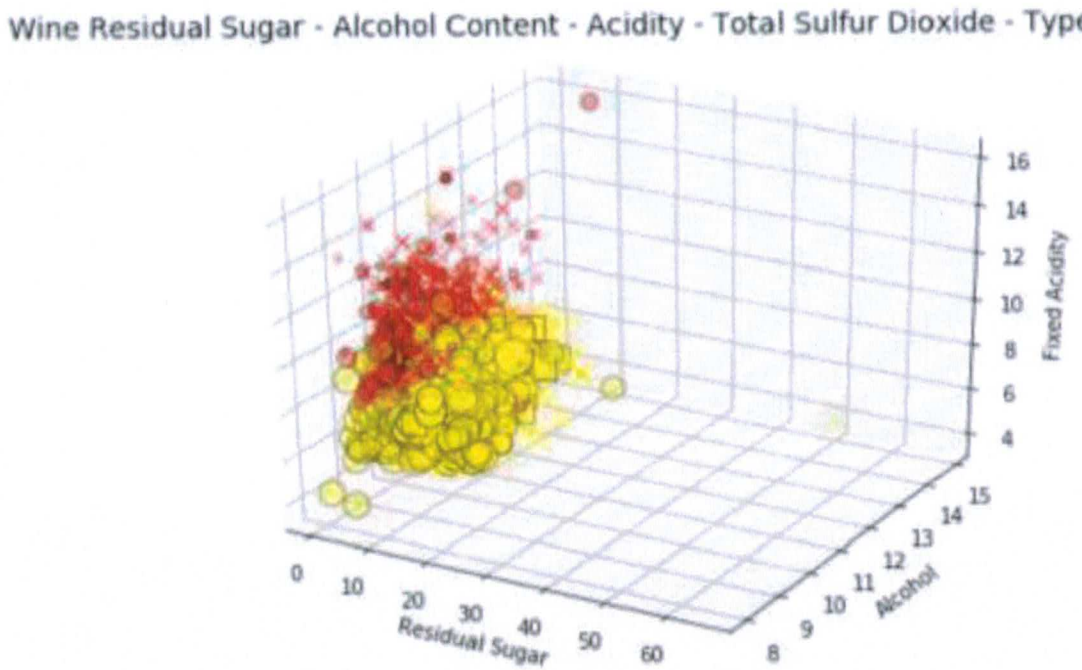
Kombinovaný graf

- Kombinuje **bodový graf a graf hustoty** - ukazuje i odhad funkce hustoty pravděpodobnosti



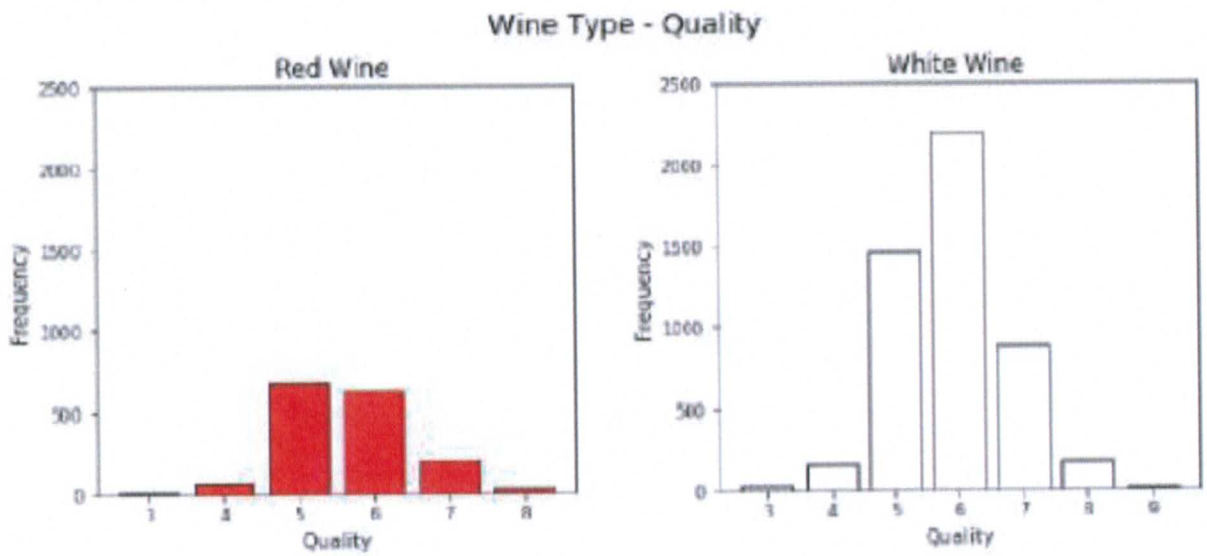
Využití vizualizačních rysů

Pro vizualizaci více atributů můžeme např. využít 3D graf, nebo využít barvu, různou velikost nebo různé tvary zobrazovaného bodu pro odlišení na základě kategorického atributu, např:



Využití soustav grafů

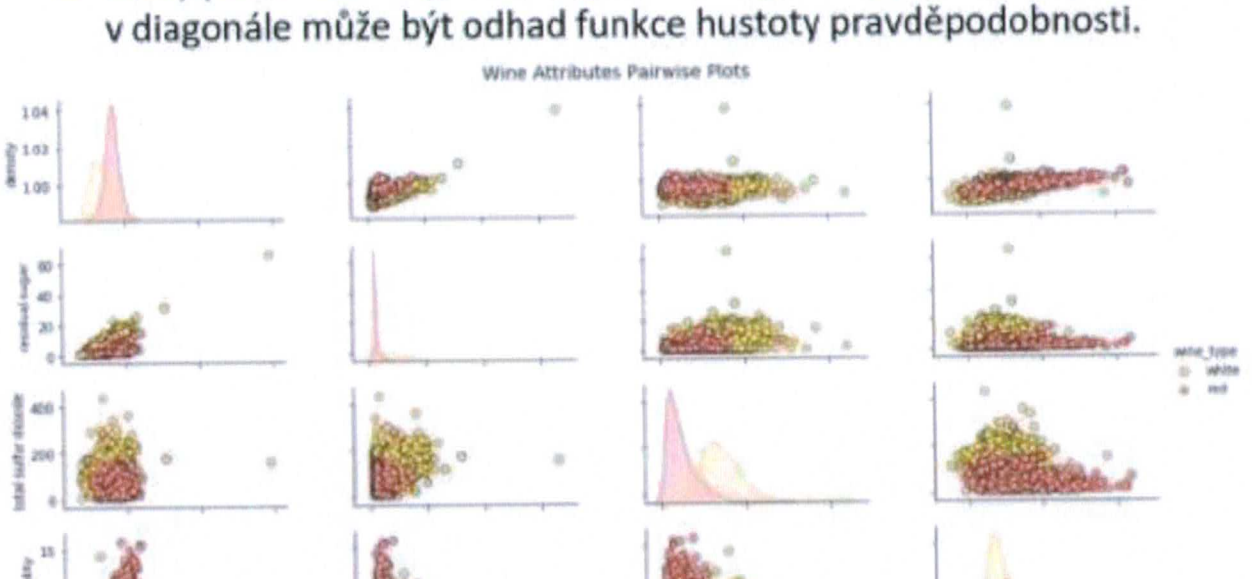
Graf můžeme rozčlenit na podgrafy (tzv. oddíly – facets), např. na základě hodnot kategorického atributu:



Matice grafů

Zobrazuje grafy pro všechny dvojice atributů

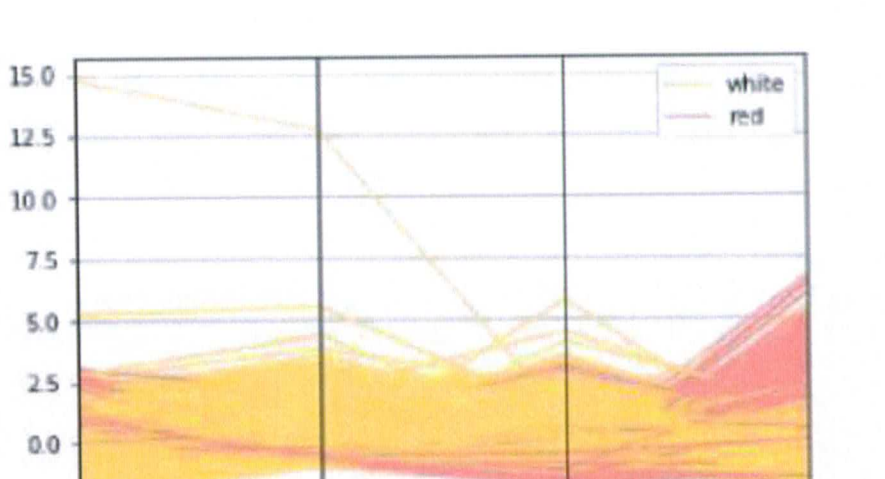
- Grafy pod a nad diagonálou symetrické (jen osy přehozené), v diagonále může být odhad funkce hustoty pravděpodobnosti.



14/2

Systém paralelních souřadnic

- Lomená čára udává hodnoty pro jeden datový objekt.
- Může pomoci odhalit shluky objektů.



5:

Korelační analýza

Korelační analýza slouží k vyjádření statistické závislosti dvou atributů. Tuto závislost můžeme vyjádřit:

- Číselně
  - pro kvantitativní atributy např. Pearsonovým nebo Spearmanovým korelačním koeficientem
  - pro kategorické atributy testem dobré shody nebo Spearmanovým korelačním koeficientem pořadí
- Vizuálně – např. pomocí bodového grafu nebo matice grafů

potřebujeme ordinální kategor. atribut

lineární závislost

- Pearsonův korelační koeficient atributů *A*, *B*

$$r_{A,B} = \frac{\sum_i (a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_i (a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

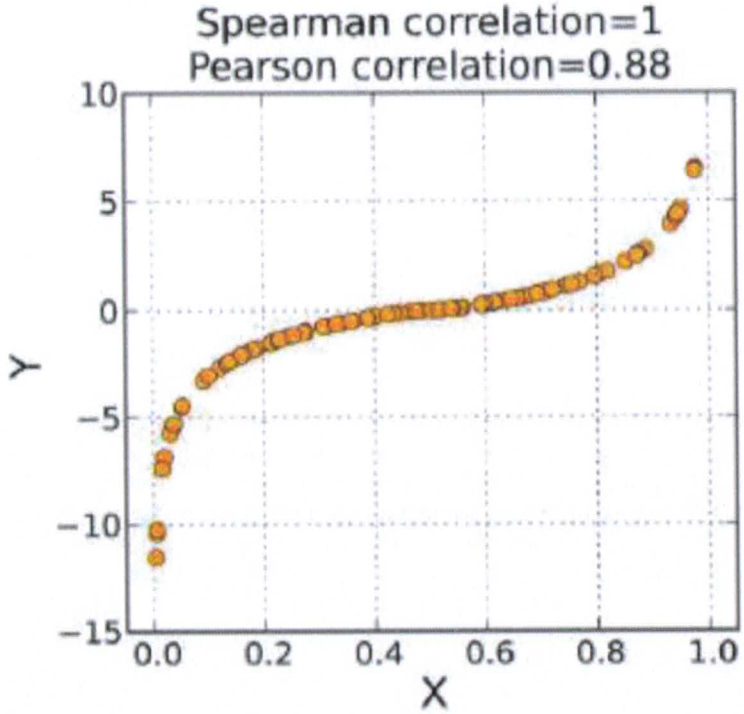
  - kde n je počet n-tic, $\bar{A}$ a $\bar{B}$ průměr atributů *A* a *B*, σ_A a σ_B jsou směrodatné odchylky *A* a *B* a a_i, b_i jsou hodnoty *A* a *B* v i- té n-tici.
  - Je-li $r_{A,B} > 0$, jsou *A* a *B* pozitivně korelované. Čím větší je hodnota koeficientu, tím větší korelace.
  - Je-li $r_{A,B} = 0$: jsou atributy nezávislé.
  - Je-li $r_{A,B} < 0$: jsou atributy negativně korelované.
  - Udává sílu a směr lineárního vztahu dvou atributů.
  - Rozsah hodnot <-1, 1>.
- Spearmanův koeficient korelace pořadí
  - Založený na pořadí uspořádaných hodnot dvou atributů
  - Vyjadřuje sílu a směr monotónního vztahu
  - Nejsou-li duplicitní hodnoty, určí se pro atributy A a B jako

$$\rho_{A,B} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

  - kde *n* je počet n-tic a $d_i = poradi(A_i) - poradi(B_i)$ je rozdíl pořadových čísel (indexů) hodnot atributů A a B v i-té n-tici.
  - Jsou-li duplicity, určí se jako Pearsonův koeficient počítaný pro pořadová čísla.
  - Opět rozsah <-1, 1>, interpretace analogická Pearsonovu, ale nejde o lineární, nýbrž monotónní korelaci.

16/2

- Lineární vs. monotónní korelace



Test dobré shody

- χ^2 - test dobré shody: H0 – nezávislost atributů

$$\chi^2 = \sum_i \sum_j \frac{(o_{i,j} - e_{ij})^2}{e_{ij}}$$

  - kde o_{ij} je pozorovaný a e_{ij} očekávaný počet výskytů (četnosti) kombinace hodnot a_i, b_j v n-ticích
  - Čím větší je hodnota χ^2 value, tím je vyšší pravděpodobnost, že jsou atributy korelované
  - K hodnotě χ^2 nejvíce přispívají ty počty kombinace hodnot a_i a b_j, které se hodně liší od očekávaných.

Určit korelaci binárních atributů *Hraje_šachy* a *Preferuje_sci-fi*.
Kontingenční tabulka:

pozorovaná

očekávaná

	Hraje šachy	Nehraje šachy	Četnost
Preferuje sci-fi	250(90)	200(360)	450
Nepreferuje sci-fi	50(210)	1000(840)	1050
Četnost	300	1200	**1500**

Výpočet χ^2 :

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

Porovnání s tabulkovou hodnotou pro stupeň volnosti $(n_A\text{-}1)*(n_B\text{-}1)$ (u nás 1) a zvolenou hladinu významnosti (např. 0.001): 10.8

→ atributy **jsou korelované**

59

Předzpracování dat

[Toto je nejspíše nad rámec otázky, ale bralo se v UPA] Dalším krokem v modelu CRISP-DM po pochopení dat je příprava dat, jejímž cílem je připravit data pro dolovací algoritmus v co nejvyšší kvalitě. Data by měla být:

- přesná
- konzistentní
- úplná (nechybí záznamy, nechybí jednotlivé atributy)

Oproti tomu reálná data bývají mnohdy neúplná, zašuměná a nekonzistentní, případně jich bývá málo. V rámci přípravy dat typicky chceme pro vstupní data provést následující:

1. výběr dat – jaká data použijeme (chceme využít kvalitní data relevantní k dolování)
2. čištění dat – zajištění kvality dat (vyřešení chybějících hodnot, zašumění, nekonzistence)
3. integrace dat – kombinace dat z několika zdrojů
4. úprava datové sady – redukce dimenzionality, počtu záznamů, řešení nevyváženosti
5. transformace – úprava dat do podoby vhodné pro dolování, např. normalizace, diskretizace

18/21

Čištění dat ke účelem zvýšení kvality dat

Chybějící hodnoty

Mohou být způsobeny různými příčinami, např. chybou zařízení, chybou zadávání, … Můžeme v zásadě ošetřit dvěma způsoby:

- ignorováním celého datového objektu
- doplnění hodnot
  - ručně
  - automaticky – nahrazení např. střední hodnotou nebo nejpravděpodobnější hodnotou

Zašuměná data

Metody odstranění šumu:

- plnění (binning)
  - data rozdělíme do košů stejné délky (stejně dlouhé intervaly, v jednotlivých koších může být různý počet položek) nebo hloubky (v každém koši přibližně stejný počet položek) a jednotlivé koše vyhladíme nahrazením nějakou střední hodnotou (průměrem, mediánem, …) nebo bližší hraniční hodnotou

Příklad: Uspořádaná data podle jednotkové ceny:

4, 9, 15, 21, 24, 24, 24, 26, 27, 28, 29, 34

1.) Rozčlenění do košů stejné hloubky:
- Koš 1: 4, 9, 15, 21
- Koš 2: 24, 24, 24, 26
- Koš 3: 27, 28, 29, 34

2a.) Vyhlazení průměrem koše:
- Koš 1: 12, 12, 12, 12
- Koš 2: 25, 25, 25, 25
- Koš 3: 30, 30, 30, 30

2b) Vyhlazení hraničními hodnotami:
- Koš 1: 4, 4, 21, 21
- Koš 2: 24, 24, 24, 26
- Koš 3: 27, 27, 27, 34

- regrese
  - hodnota několika atributů je použita k predikci hodnot jiného atributu
- analýza odlehlých objektů – např. pomocí shlukování

19/2

Úprava datové sady

Redukce dimenzionality

S rostoucím počtem atributů (dimenzí) často roste složitost algoritmů, proto pro některé algoritmy chceme dimenzionalitu snížit. Jsou v zásadě tři přístupy, jak snížit dimenzionalitu:

- konstrukce rysů – vytvoření nového atributu z existujících (např. celková útrata namísto útrata za ubytování, útrata za stravu)
- extrakce rysů – hledáme deskriptory, které nejlépe popisují zdrojová data. Typicky se jedná o transformaci nového prostoru atributů (kterých je méně než původních), např. metoda Principal Component Analysis (PCA)
- výběr rysů – hledání nejmenší podmnožiny atributů, která je užitečná s ohledem na zpracování (nepotřebnou informaci odstraníme)

Řešení nevyváženosti dat

Klasifikační algoritmy často předpokládají, že třídy budou v trénovacích datech vyváženy, ovšem reálná data takový charakter často nemají (např. negativních covid testů bude více než pozitivních). Máme několik přístupů jak tento problém řešit:

- podvzorkovat majoritní třídu (náhodné podvzorkování nebo heuristické metody, co se snaží vyloučit méně důležitá data)
- nadvzorkovat minoritní třídu (zduplikovat data)
- komplikovanější přístupy, např. metoda SMOTE vychází z principu, že data blízko minoritní třídy budou nejspíše také patřit do minoritní třídy. Snažíme se tedy syntetizovat nová data minoritní třídy následujícím algoritmem:

```
FOR EACH instanci minoritní třídy c̄ DO
    Najdi jeho K nejbližších sousedů
    Náhodně vyber jednoho z nich k̄
    Vytvoř novou instanci minoritní třídy n̄ jako
        n̄ = c̄ + (k̄ − c̄) * rand(0, 1)
```

Transformace dat

Normalizace

Cílem normalizace je převést data do určitého rozsahu, např. min-max normalizace nejprve převede hodnotu do rozsahu <0, 1> a poté případně přeškáluje hodnoty na nové minimum a maximum:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Případně pokud data vychází z normálního rozložení, můžeme využít Z-score normalizaci:

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Diskretizace

Některé algoritmy nepracují s numerickými atributy a fungují správně pouze s kategorickými – potřebujeme tedy převést spojité hodnoty na diskrétní. K diskretizaci se může využívat např. binning a přiřazení návěští jednotlivým intervalům, viz výše.

Kódování

Některé algoritmy naopak pracují pouze s numerickými atributy a je potřeba tedy převést kategorické atributy. K tomu se využívá kódování, které je možno provést např. následujícími způsoby:

- ručně – typicky pro ordinální atributy, kde je známé pořadí
- automaticky převod řetězce na číslo
- binární – převedeme řetězec na celé číslo, z něj uděláme binární kódování a pro každý bit zavedeme sloupec

Pokud v textu najdete chybu, nebudete něčemu rozumět nebo budete mít dojem, že by bylo vhodné něco doplnit, kontaktujte na discordu uživatele Fifinas.