

Porozumění datům



Ivana Burgetová

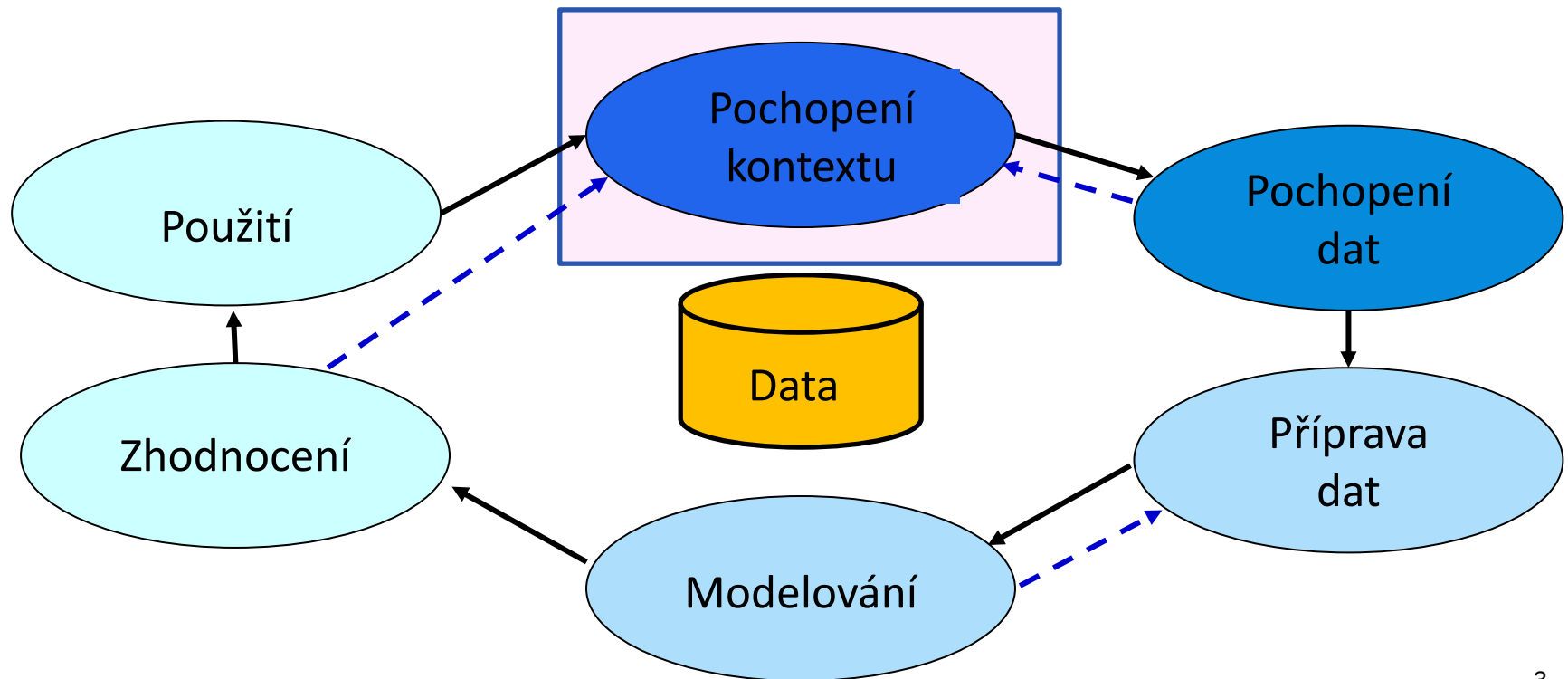
FIT VUT v Brně

Obsah

- Krok porozumění datům
- Datová sada (dataset)
- Popisné charakteristiky dat
- Vizualizační techniky
- Korelační analýza

Krok pochopení kontextu

- Model životního cyklu analytického projektu CRISP

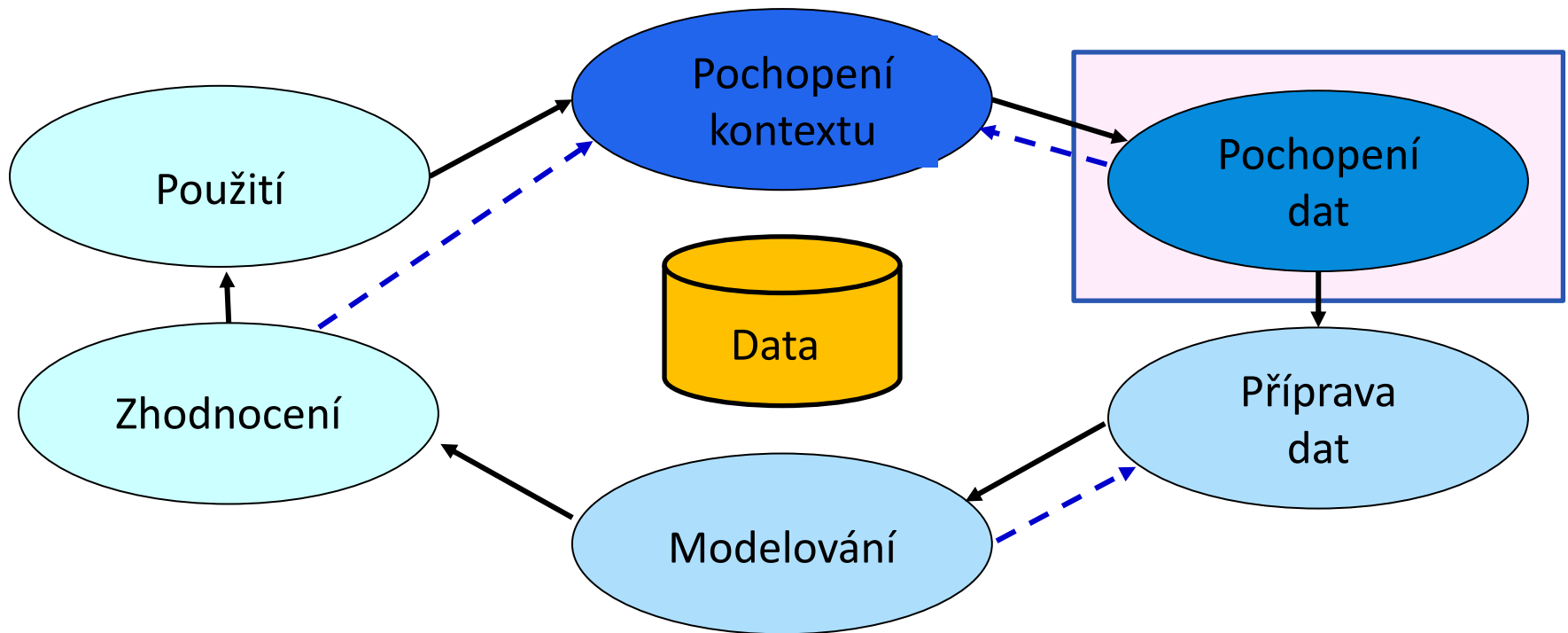


Krok pochopení kontextu

- **Cíl:** Dozvědět se co nejvíce o kontextu řešeného problému (pochopení dané problematiky, *business understanding*).
- **Zahrnuje:**
 1. Pochopení a určení **cílů projektu** a požadavků z hlediska cílů organizace/výzkumu/... + **kritéria úspěšnosti** projektu
 - *Příklad:* Jak více upoutat zákazníka na našich webových stránkách? (doporučování, personalizace)
 2. **Zhodnocení situace** – dostupné zdroje pro řešení (lidé + jejich znalosti, **data** – jsou k dispozici?, omezení, ...)
 3. Určení **cílů z hlediska DM** - převedení znalosti cílů do podoby definice DM problému + **kritéria a metody hodnocení**
 - *Příklad:* Na základě analýzy historických dat najdeme skupiny produktů kupovaných současně. (analýza nákupního košíku), zvýšení času o 5%.
 4. Vytvoření **projektového plánu**
(ekonomický rozbor, plánování jednotlivých fází, analýza rizik, ...)

Krok porozumění datům

- Model životního cyklu analytického projektu CRISP



Krok porozumění datům

- **Cíl:** Dozvědět se o datech co nejvíce (*data understanding*).
- **Předpoklady** (z pohledu dat):
 - Chápeme problematiku
 - Máme stanovený cíl projektu a kritéria úspěšnosti
 - Víme, jaká data jsou/budou k dispozici
 - Víme, o jakou základní DM úlohu (příp. několik) půjde
- **Vstup:**
 - Dostupné datové zdroje a dokumentace k nim
- **Výstup:**
 - Podklady pro vytvoření počáteční datové sady nebo přímo vytvoření
 - Detailní informace o vlastnostech dat (popisné charakteristiky + grafy)

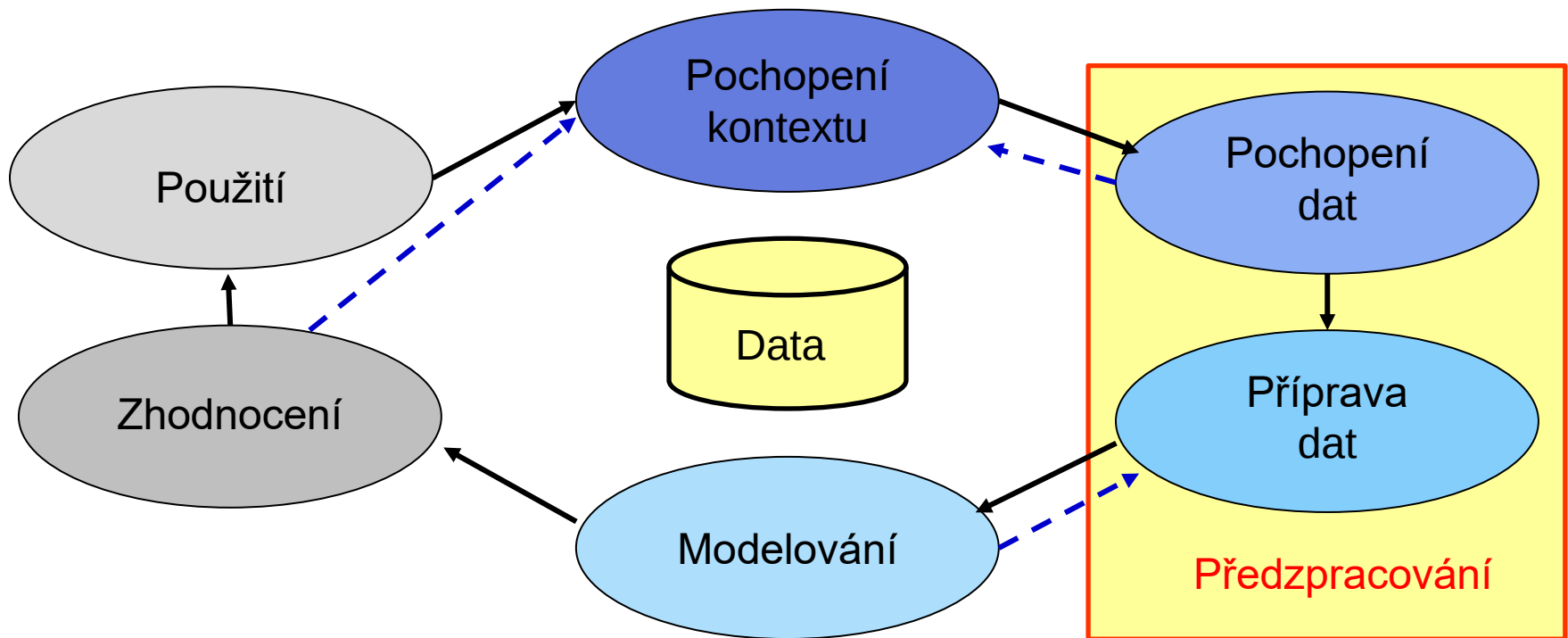
Krok porozumění datům

■ Obvykle zahrnuje kroky:

1. **Rozpracování informace** k dostupným/potřebným datům z kroku pochopení kontextu.
 - jaká data máme – význam, dostupnost, cena
 - věrohodnost dat
 - potřeba pořízení dalších dat – cena
 - při použití více zdrojů – kdy a jak integrovat
2. **Popis dat**
 - struktura a význam, formát, množství
3. **Prozkoumání dat (explorační analýza)**
 - popisné statistické charakteristiky, grafy, korelace, ...
4. **Zhodnocení kvality dat**
 - úplnost – chybějící vlastnosti, chybějící hodnoty, šum, ...

Krok porozumění datům

- Model životního cyklu analytického projektu CRISP

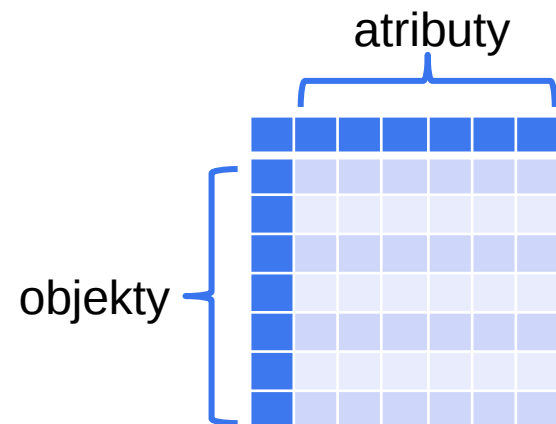


Obsah

- Krok porozumění datům
- Datová sada (dataset)
- Popisné charakteristiky dat
- Vizualizační techniky
- Korelační analýza

Datová sada

- **Datová sada/dataset** – kolekce dat vybraných k dolování
 - Rozdělení – viz rozdělení zdrojů dat (tradiční, netradiční (pokročilá data))
 - Jsou tvořeny datovými objekty
- Datové objekty tradičních datových sad popsány *atributy*
- Datová sada tradičních dat má podobu tabulky (flat file)
 - Řádky tabulky → datové objekty
 - Sloupce → atributy



Datový objekt

- Struktura typicky reprezentující entity reálného světa
 - obvykle chápeme význam jednotlivých sloupců:
 - *Příklad:* Datová sada prodeje: zákazníci, skladové položky, prodeje
 - ale ne vždy:
 - *Příklad:* vektor reprezentující slovo/obrázek (embedding)
- Nazývají se také *vzorky (samples), příklady (examples), instance, datové body, objekty, n-tice*

Datový objekt

- **Atribut** (také dimenze, rys/příznak (feature), proměnná)
 - část struktury datového objektu reprezentující nějakou jeho vlastnost/rys.
 - *Příklad:*
 - ID zákazníka, jméno, adresa
 - četnost výskytu termu v dokumentu, histogram barev oblasti
 - Typy atributů
 - Kategorický
 - Numerický/kvantitativní

Typy atributů

- **Kategorický** – řetězec nebo celočíselný (kvalitativní)
 - **Binární** – nominální se dvěma stavy
 - **Symetrický binární**: oba stavy stejně důležité
 - *Příklad*: pohlaví $\in$ {muž, žena}
 - **Asymetrický binární**: stavy nemají stejnou důležitost
 - *Příklad*: test $\in$ {pozitivní, negativní}
 - **Nominální**: kategorie, stavy, *jména věcí*
 - *Příklad*: barva $\in$ {černá, modrá, bílá}
 - **Ordinální** – je definováno uspořádání, ale rozdíl dvou sousedních hodnot není definován
 - *Příklad*: vzdělání $\in$ {základní, střední všeobecné, střední odborné, vysokoškolské}

Typy atributů

□ Kvantitativní – celočíselný nebo reálný

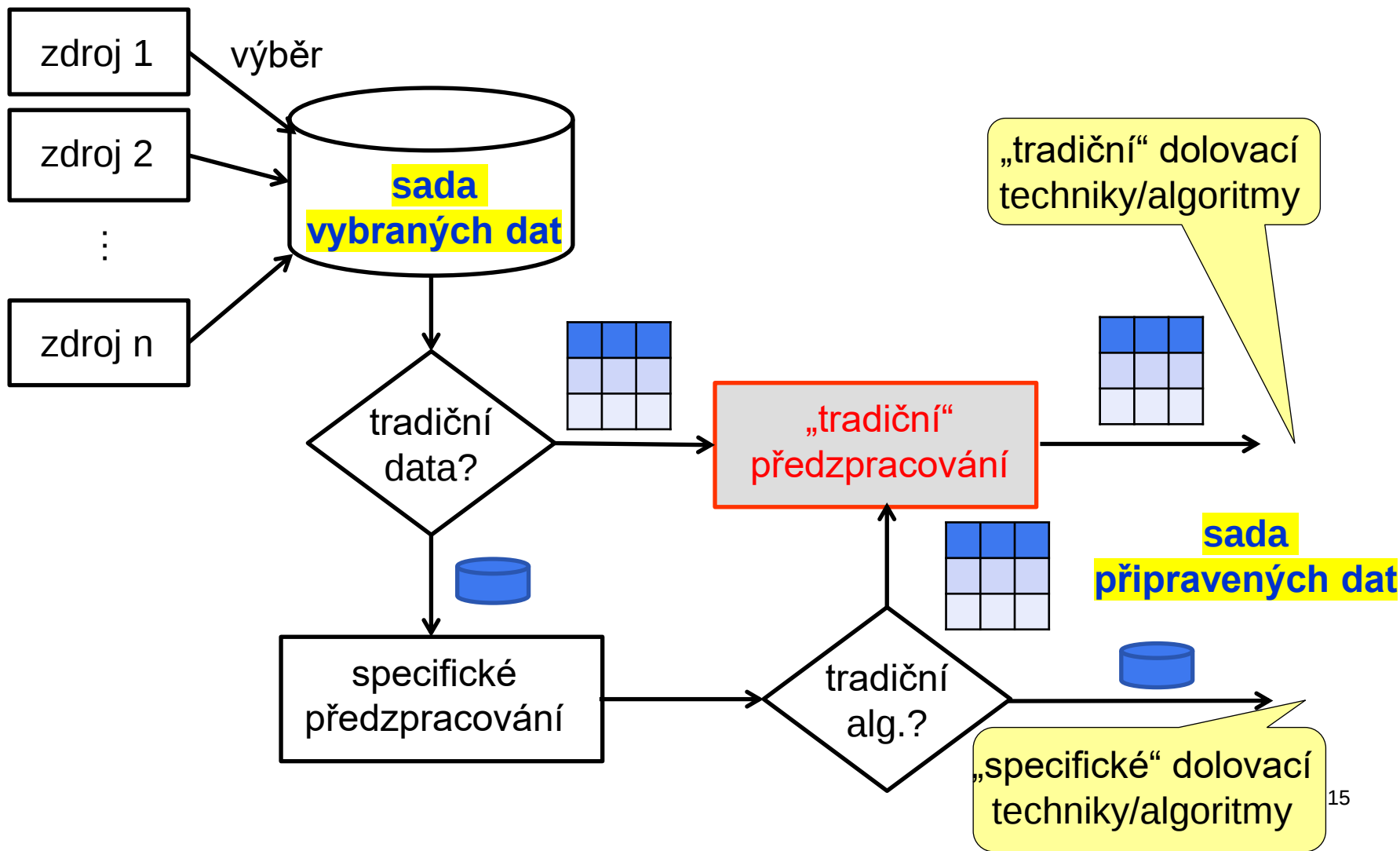
■ Intervalový

- hodnoty jsou uspořádatelné
- spojitě měřítko dělené přibližně lineárně, jsme schopni určit rozdíl (tj. navíc od ordinálního), podíl je problematický
- není vztažný/nulový bod
 - *Příklad:* teplota ve °C, datum

■ Poměrový

- implicitní nulový bod
- oproti intervalovým jsme schopni určit i podíl dvou hodnot
- jen kladné hodnoty
 - *Příklad:* teplota v K, délka, počet dětí, částka

Datová sada



Obsah

- Krok porozumění datům
- Datová sada (dataset)
- Popisné charakteristiky dat
- Vizualizační techniky
- Korelační analýza

Popisné charakteristiky dat

□ Motivace:

- Pochopení dat – zejména **rozložení hodnot** atributů

□ Datový soubor obsahující n objektů (pozorování)

- vzorek celé (příp. nekonečné) populace pozorování hodnot náhodné veličiny
- zjistíme míry **náhodného výběru** → míry **populace**

□ Kategorie statistických měr datového souboru

- **míry polohy** – určují *střed* dat, případně další *body* z hlediska rozložení hodnot
- **míry variability** – určují rozptýlenost hodnot kolem *středu*
- další (např. šikmost, špičatost)

Popisné charakteristiky dat

- Typy měř z hlediska možnosti distribuce výpočtu
 - **Distributivní** – lze počítat distribuovaně, výsledek stejnou operací s mezivýsledky.
 - *Příklad:* počet prvků
 - **Algebraická** – výsledek algebraickou operací nad jednou nebo několika distributivními mírami.
 - *Příklad:* průměr
 - **Holistická** – lze spočítat pouze výpočtem nad celým souborem.
 - *Příklad:* medián

Míry polohy

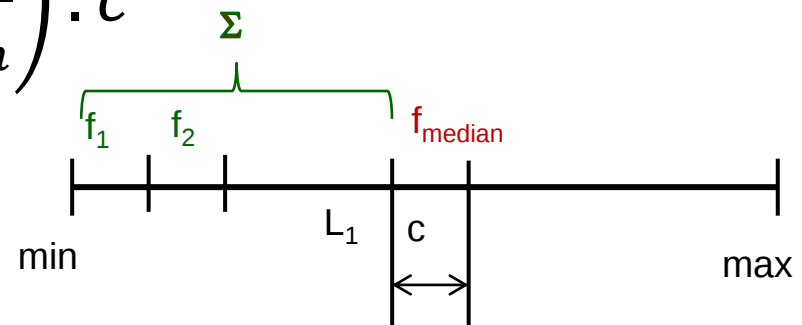
- **Aritmetický průměr** (střední hodnota): $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ $\mu = \frac{\sum x}{N}$
 - **Vážený průměr:** $\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$
 - citlivost na odlehlé hodnoty
 - **Upravený průměr (trimmed mean):** ořezání extrémních hodnot
- **Geometrický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

Míry polohy

■ 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):

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



■ Modus:

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

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

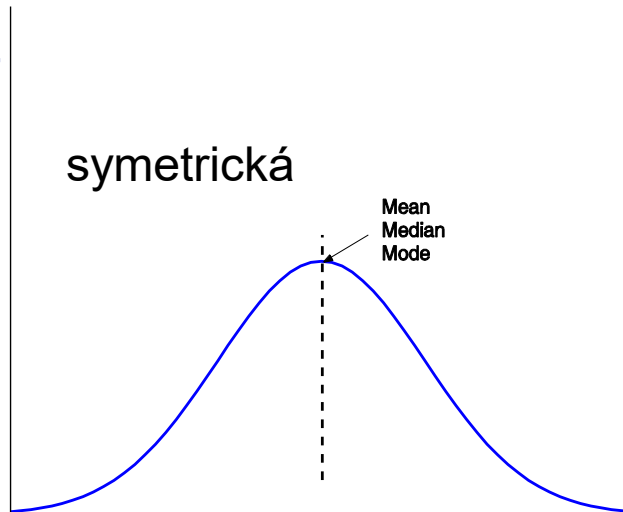
Míry polohy

□ Symetrická a vychýlená data

- Medián, modus a průměr u symetrických dat splývají

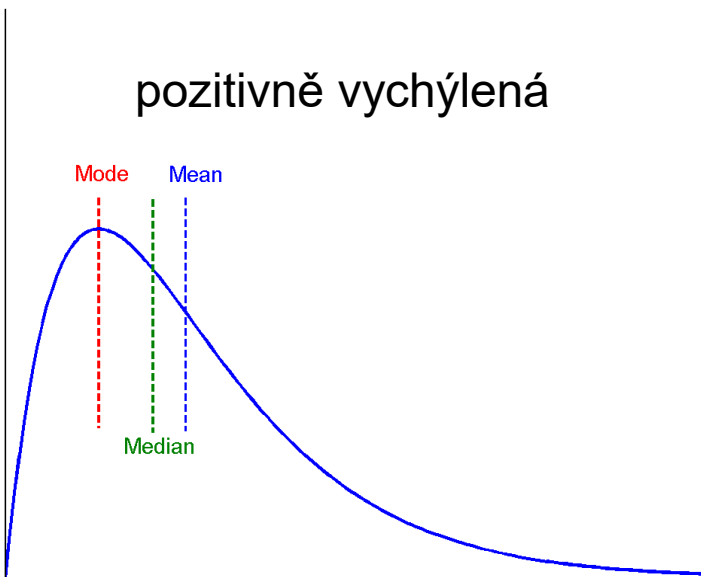
symetrická

Mean
Median
Mode



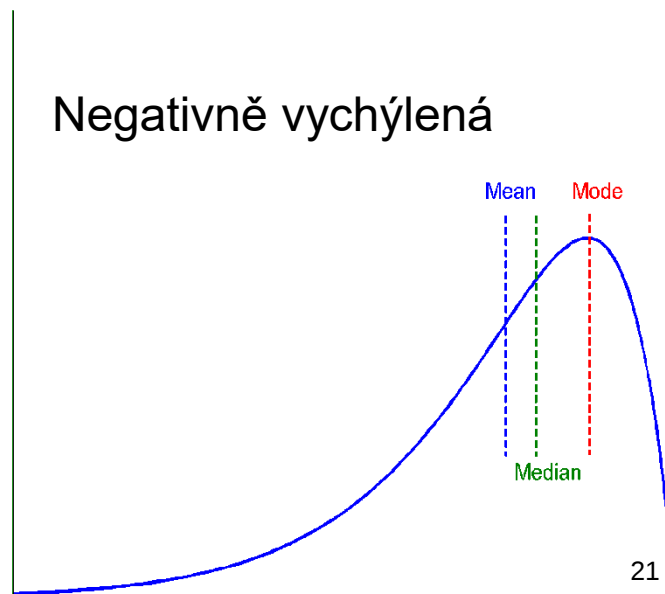
pozitivně vychýlená

Mode
Mean
Median



Negativně vychýlená

Mean
Mode
Median



Míry polohy

Příklad: Uvažujme soubor hodnot {10, 3, 1, 2, 3}

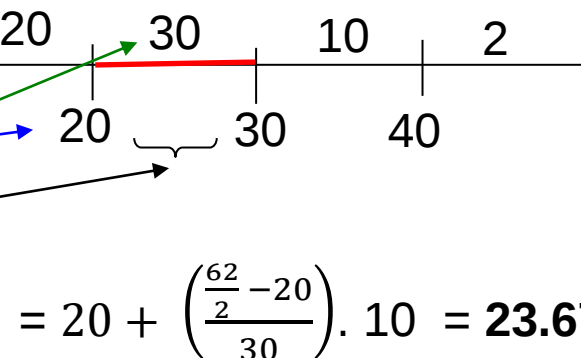
$$\bar{x} = 19/5 = 3.8, \bar{x}_G = \sqrt[5]{180} = 2.83, \bar{x}_H = 2.2, \text{medián} = 3$$

Příklad: Uvažujme znalost počtu hodnot v intervalech

interval	<20	<20, 30)	<30, 40)	≥40
četnost	20	30	10	2

Odhad mediánu?

$$\text{median} = L_1 + \left(\frac{\frac{n}{2} - \sum f_l}{f_{\text{median}}} \right) \cdot c$$



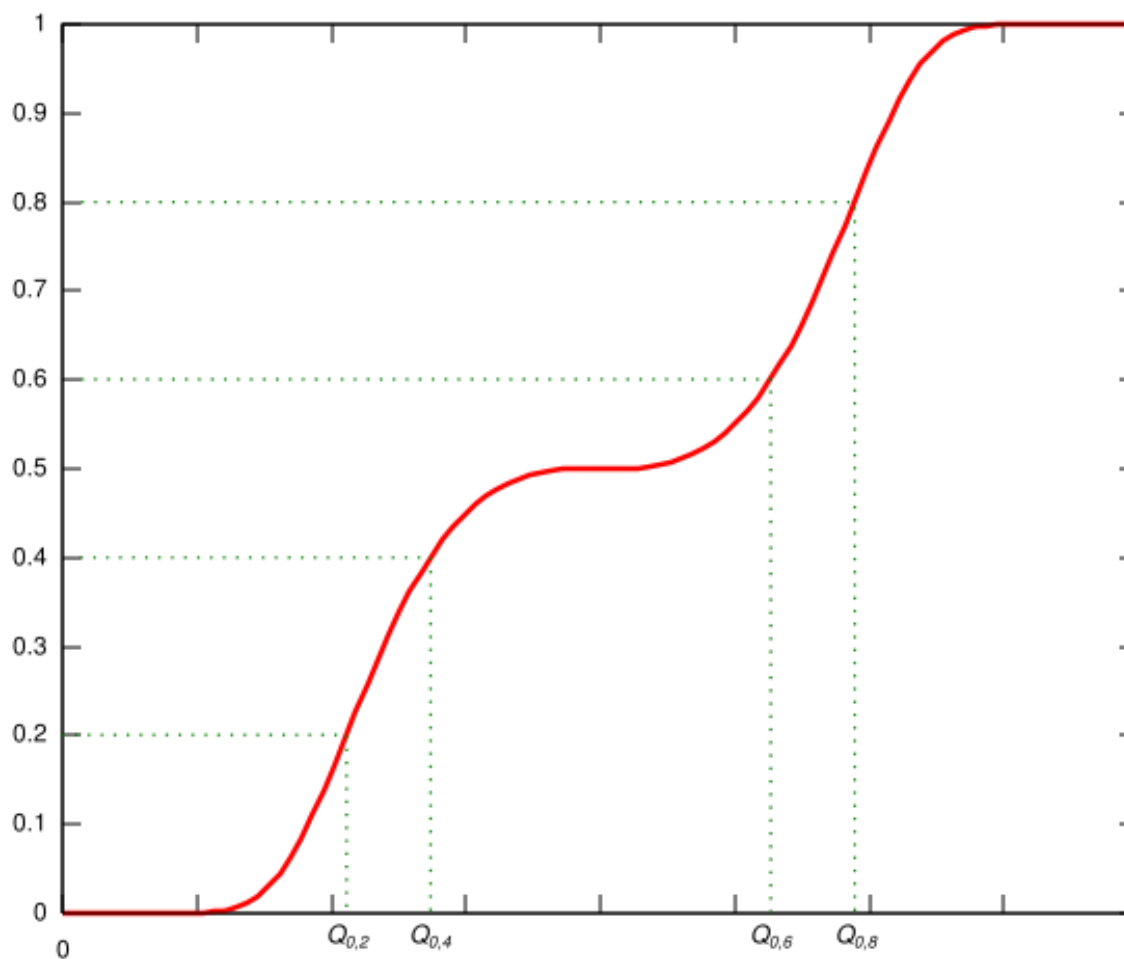
Míry polohy

□ *Kvantily:*

- Míra tvořená skupinou $q-1$ hodnot q_k :
 - q – stupeň kvantilu (počet intervalů)
 - k – pořadové číslo kvantilu
- **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 hodnot na q přibližně stejně početných intervalů
- Pro $q = 100$ – percentily
 - $q = 10$ – decily
 - $q = 5$ – kvintily
 - $q = 4$ – kvartily**
- Kvartily: q_1 (25. percentil), q_2 medián, q_3 (75. percentil)

Míry polohy

■ *Kvantily (pokrač.)*



$q = 5$

Míry variability dat

□ **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)

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

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

□ **Směrodatná odchylka**

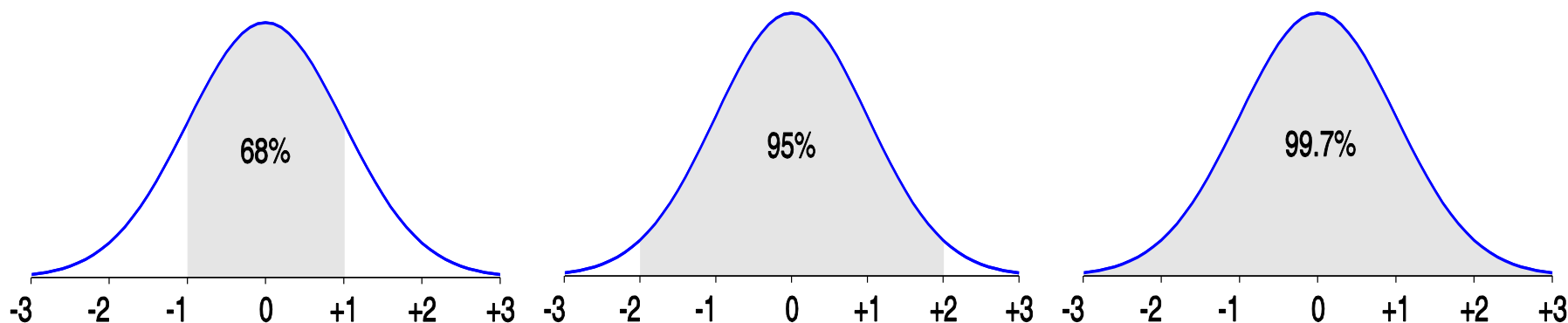
$$s = \sqrt{s^2} \qquad \sigma = \sqrt{\sigma^2}$$

Míry variability dat

□ Rozptyl normálního rozdělení dat

- pro křivku hustoty normálního rozdělení platí:

- V intervalu $\langle \mu - \sigma, \mu + \sigma \rangle$ leží asi 68% hodnot
- V intervalu $\langle \mu - 2\sigma, \mu + 2\sigma \rangle$ leží asi 95% hodnot
- V intervalu $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$ leží asi 99.7% hodnot



Míry variability dat

□ 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*)

$$MAD = \text{median}(|x_i - \text{median}|)$$

- robustní míra málo ovlivněná odlehlými hodnotami

■ (Průměrná) absolutní odchylka (*mean absolute deviation*)

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

□ Variační koeficient

$$k = \frac{s}{\bar{x}}$$

□ Sumarizace 5 čísel - složená míra

$$< \min, q_1, \text{median}, q_3, \max >$$

Obsah

- Krok porozumění datům
- Datová sada (dataset)
- Popisné charakteristiky dat
- Vizualizační techniky
- Korelační analýza

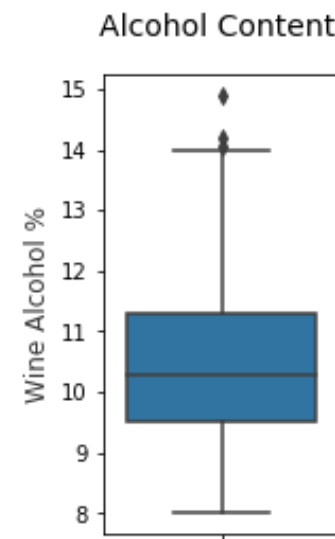
Vizualizační techniky

- Grafy pro **jeden atribut**
 - Cíl – ukázat **rozložení hodnot** atributu
 - **Kvantitativní** atributy
 - Krabicový graf
 - Histogram s rozdělením na intervaly
 - Graf hustoty
 - Houslový (violin) graf
 - Kvantilový graf
 - **Kategorické** atributy
 - Histogram

V prezentaci jsou využity obrázky grafů z [1] (prodeje zboží v prodejnách) a [4] (datová sada portugalských mladých vín (Vinho Verde)).

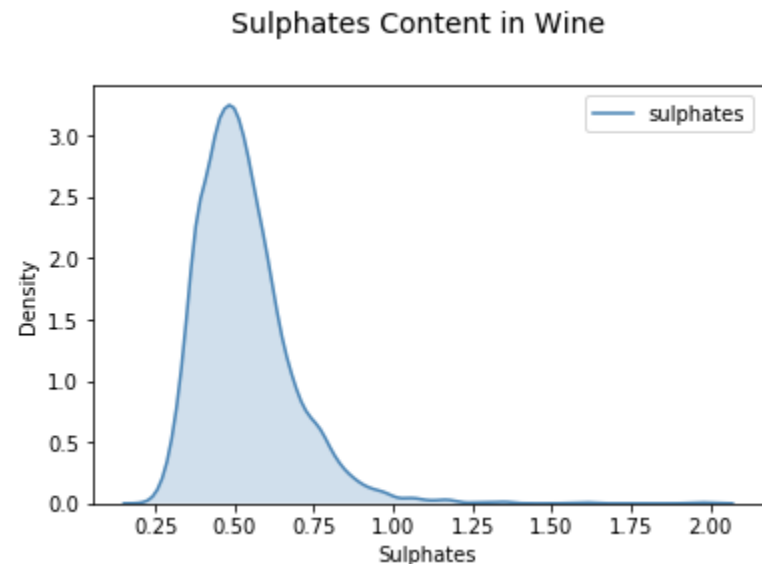
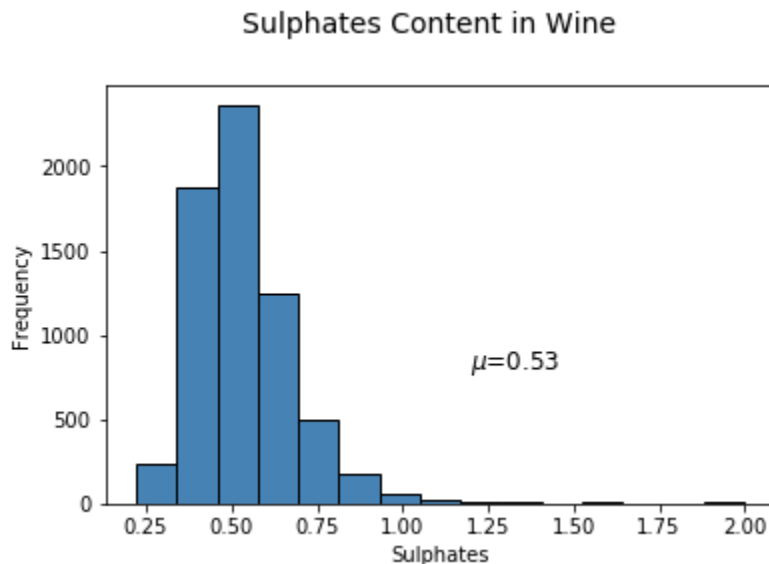
Krabicový graf

- Vizualizuje sumarizaci 5 čísel
- 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



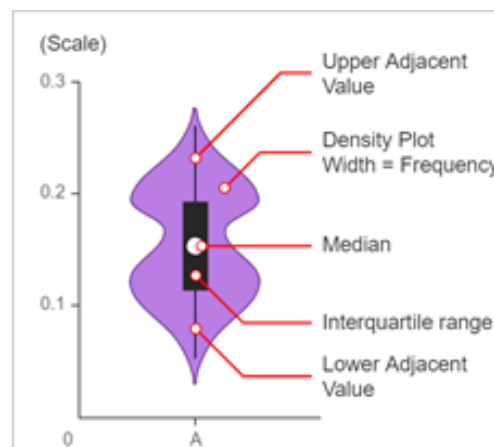
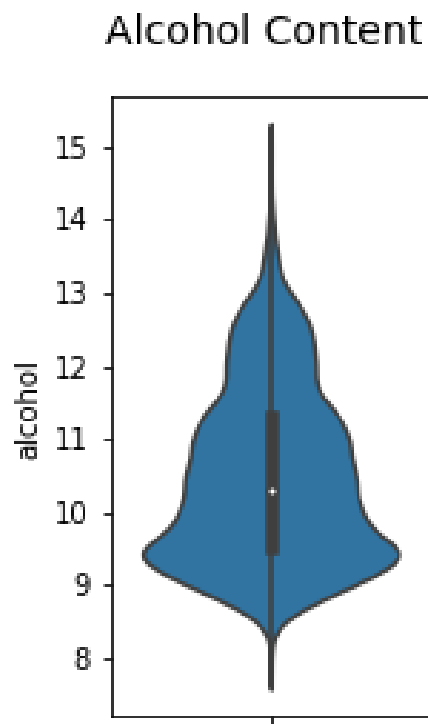
Histogram a graf hustoty

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



Houslový graf

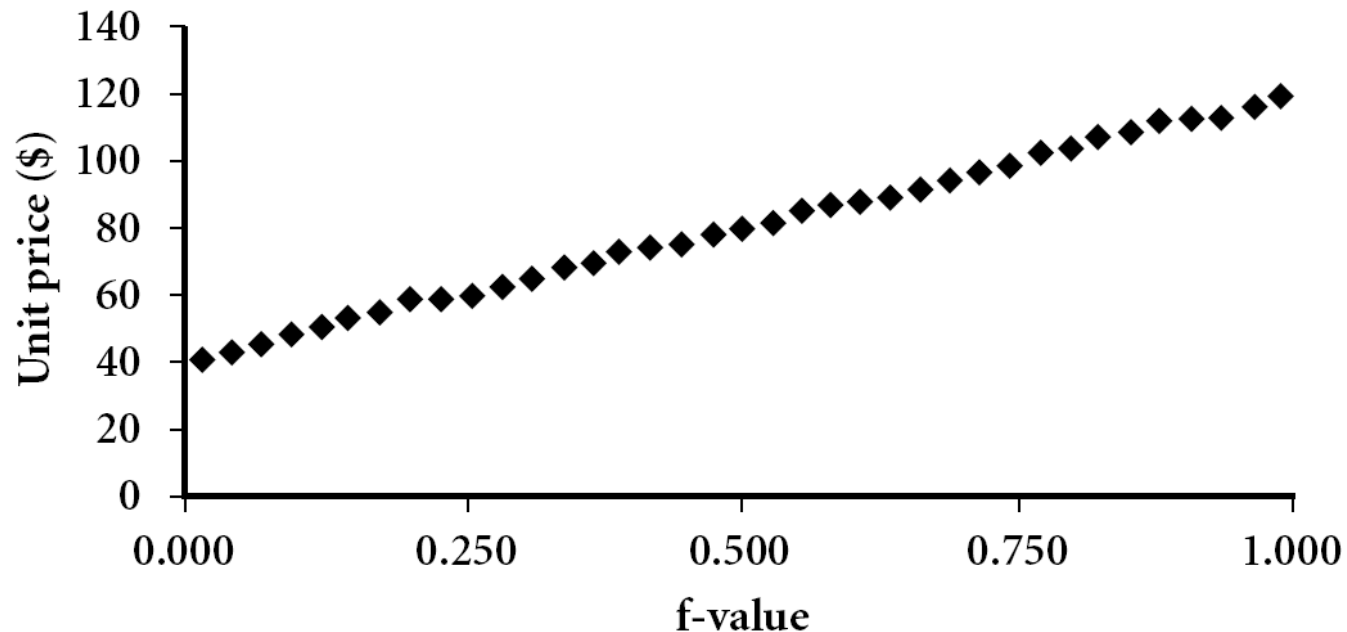
- Podobný krabicovému, ale ukazuje navíc odhad hustoty pravděpodobnosti.
 - Vhodný zejména pro multimodální data.



Zdroj [5]

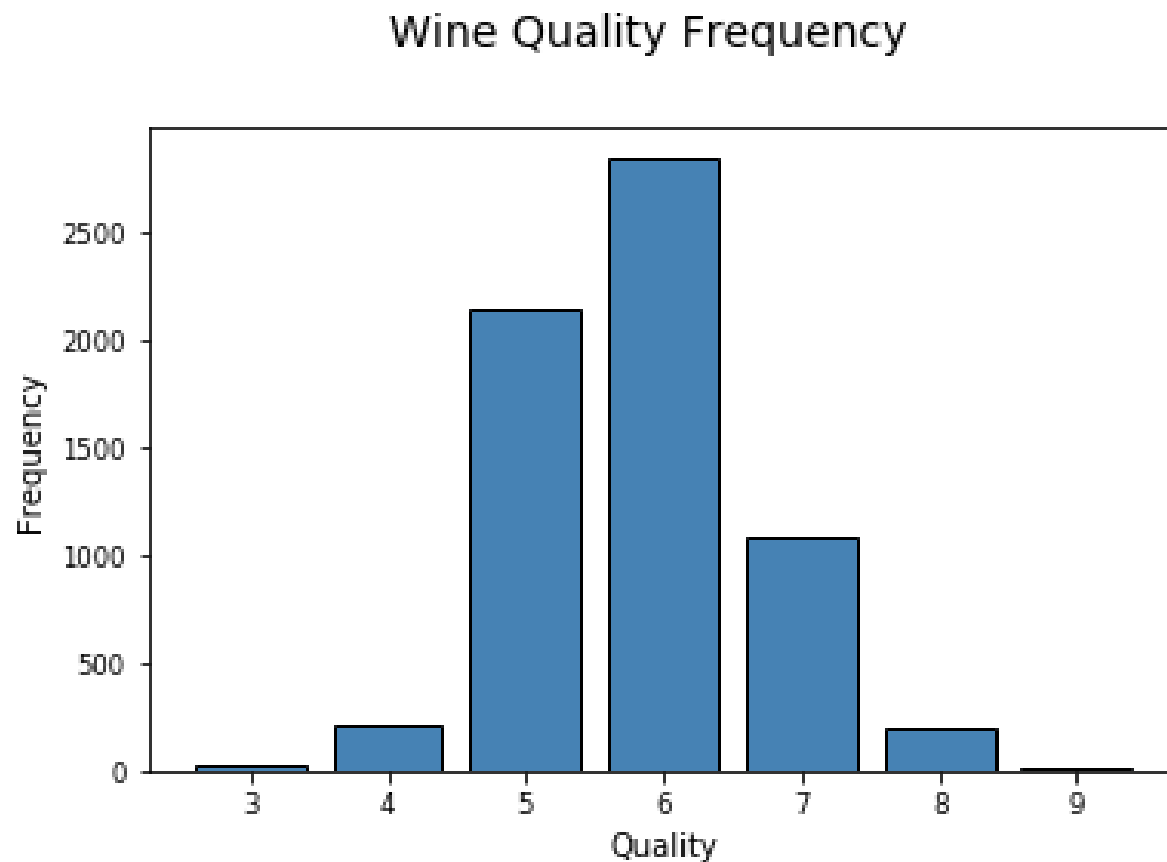
Kvantilový graf

- 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



Histogram

- Pro kategorický atribut

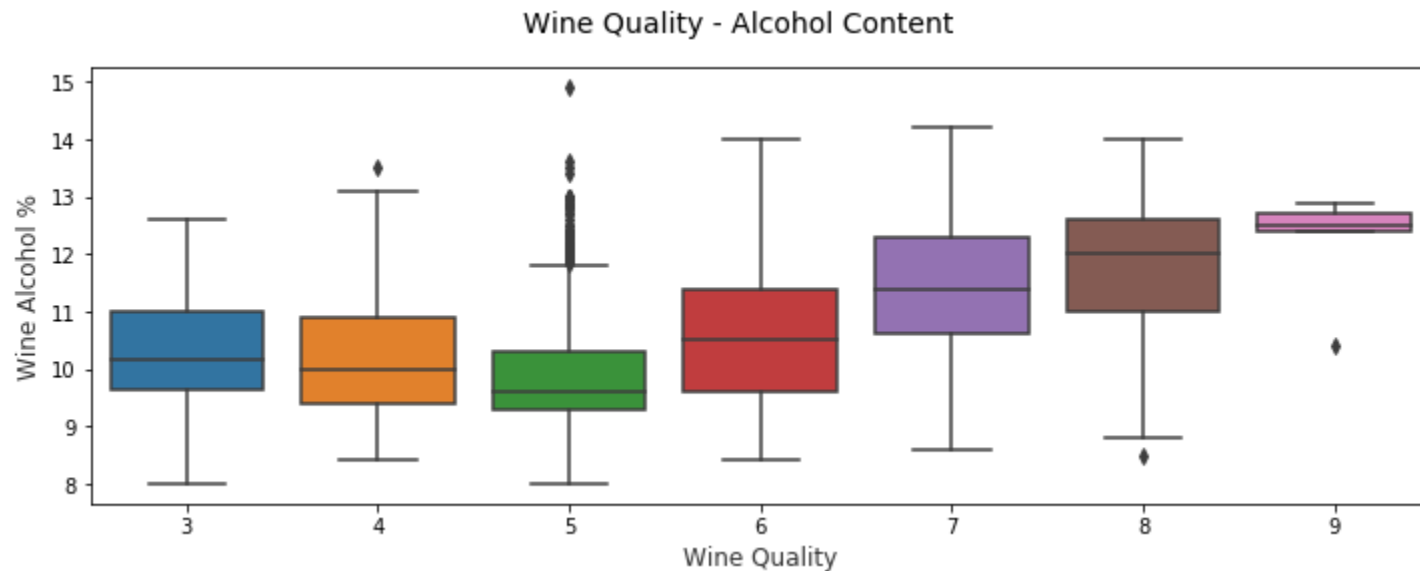


Vizualizační techniky

- Vizuální analýza **několika atributů**
- Cíl:
 - Porovnání **rozložení hodnot**
 - Odhalení potenciálních **vztahů mezi atributy** (korelace)
- Metody **rozšíření** z analýzy jednoho atributu:
 - Technika **pro jeden atribut umožňuje vizualizovat** i více atributů:
 - Krabicový/houslový graf – kvantitativní vs. kategorický atribut
 - Histogram – několik kategorických, případně kvantitativních s rozdělením na stejné intervaly, vs. četnost.
 - **Speciální typy grafů** pro dvojici atributů (vztahy) – viz dále.
 - Vizualizace dalšího atributu použitím nějakého **vizuálního rysu** – viz dále.
 - Sestavy grafů a speciální grafy – viz dále.

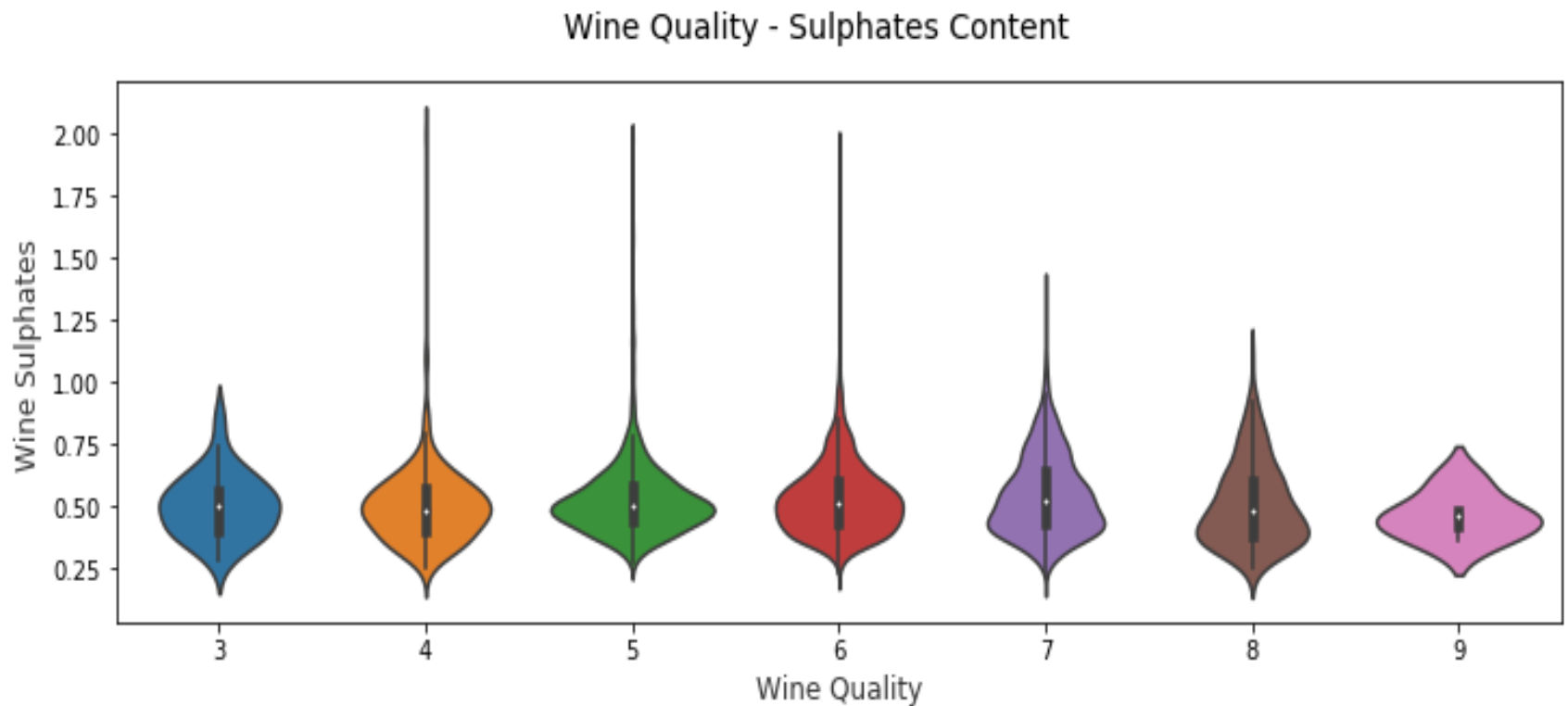
Krabicový graf

- Pro dvojici atributů - kvantitativní vs. kategorický



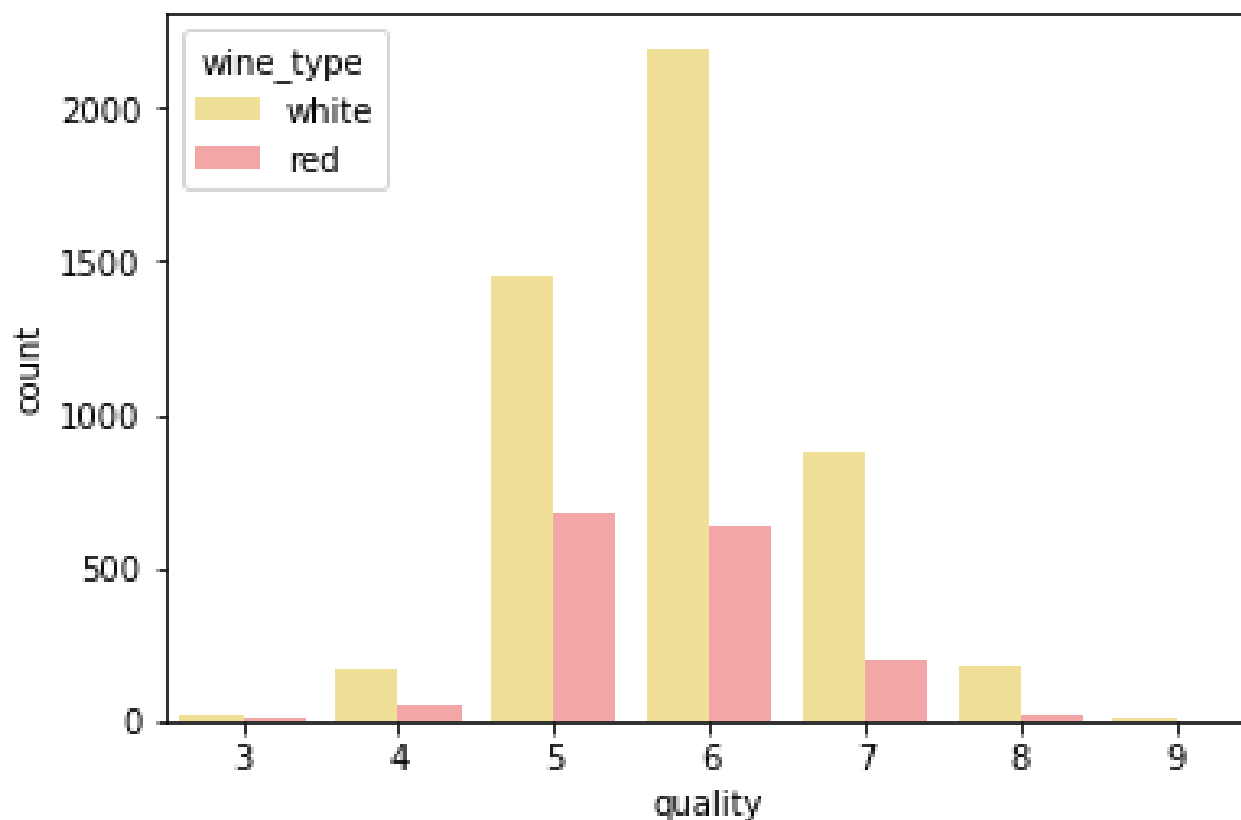
Houslový graf

- Pro dvojici atributů - kvantitativní vs. kategorický



Histogram

- Histogram/sloupcový graf pro 2 kategorické atributy

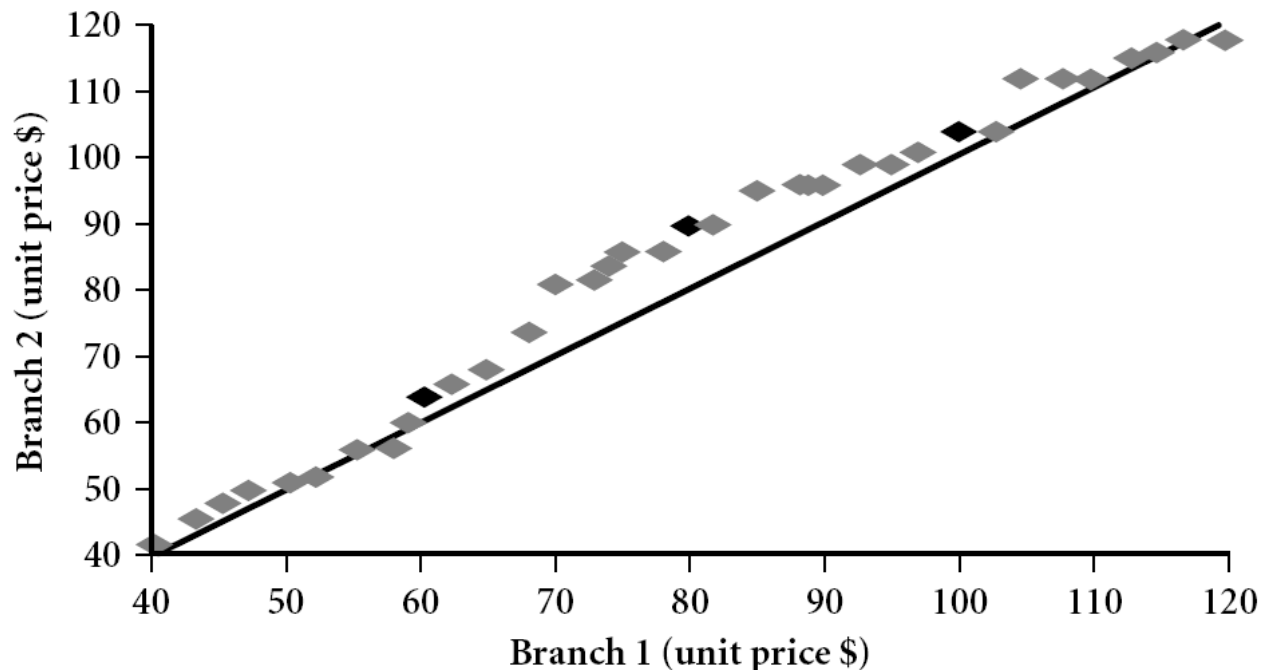


Vizualizační techniky

- Vizuální analýza **několika atributů**:
 - Metody **rozšíření** z analýzy jednoho atributu (pokrač.)
 - **Speciální typy grafů** pro dvojici atributů (vztahy):
 - Q-Q graf
 - Bodový graf, křivka lokální regrese (Loess)
 - Kombinovaný graf – bodový + hustoty

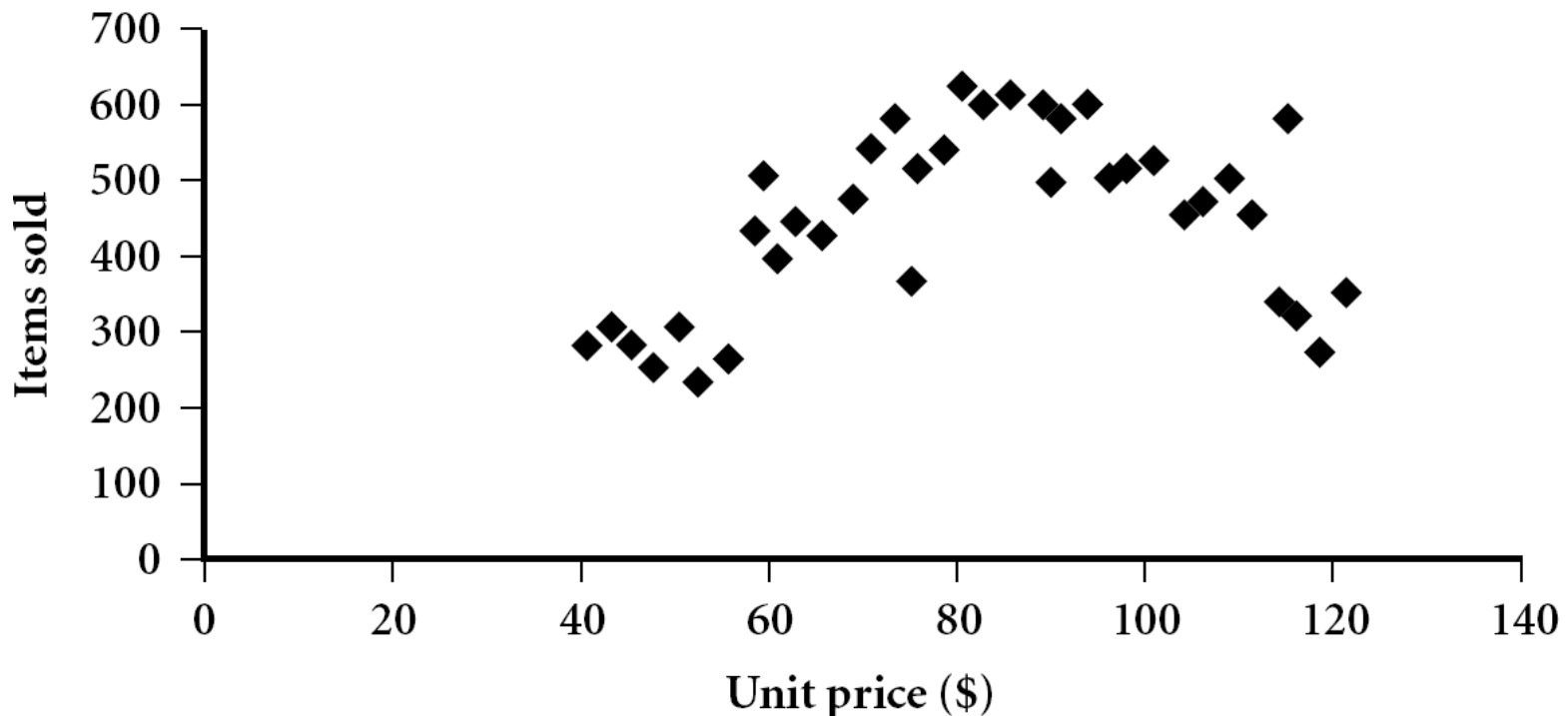
Q-Q graf

- 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



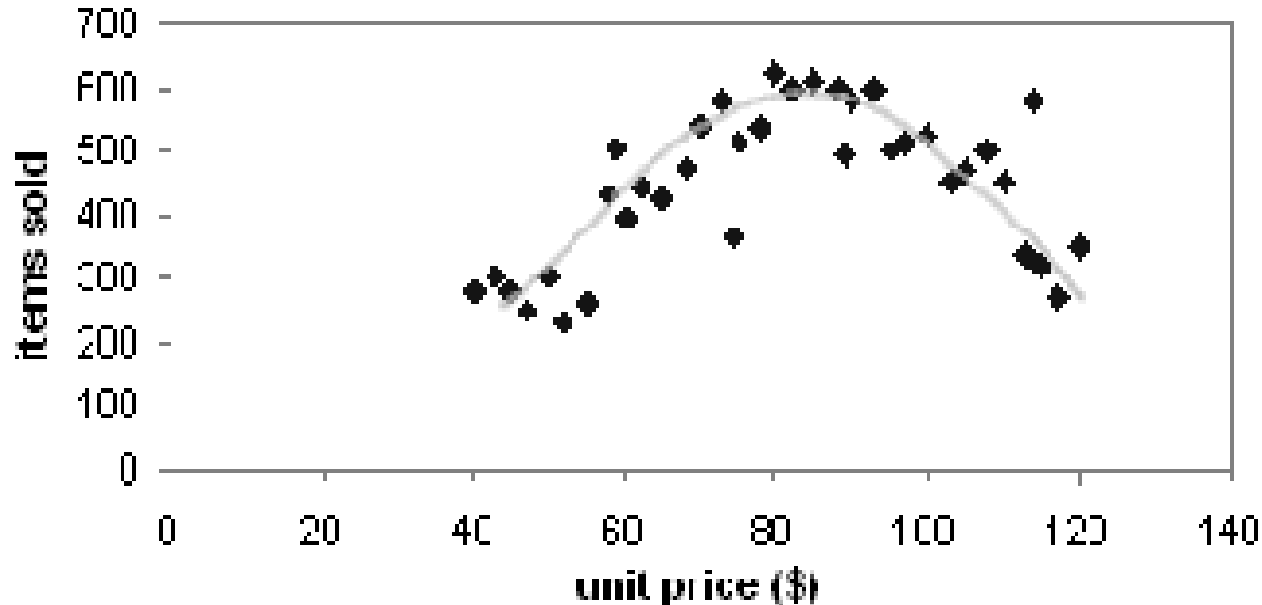
Bodový graf

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



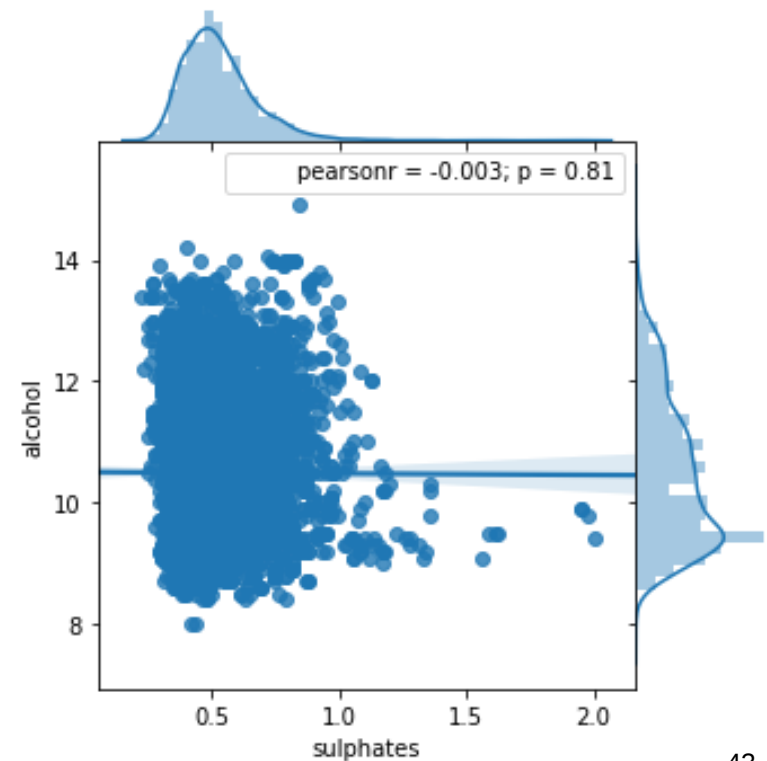
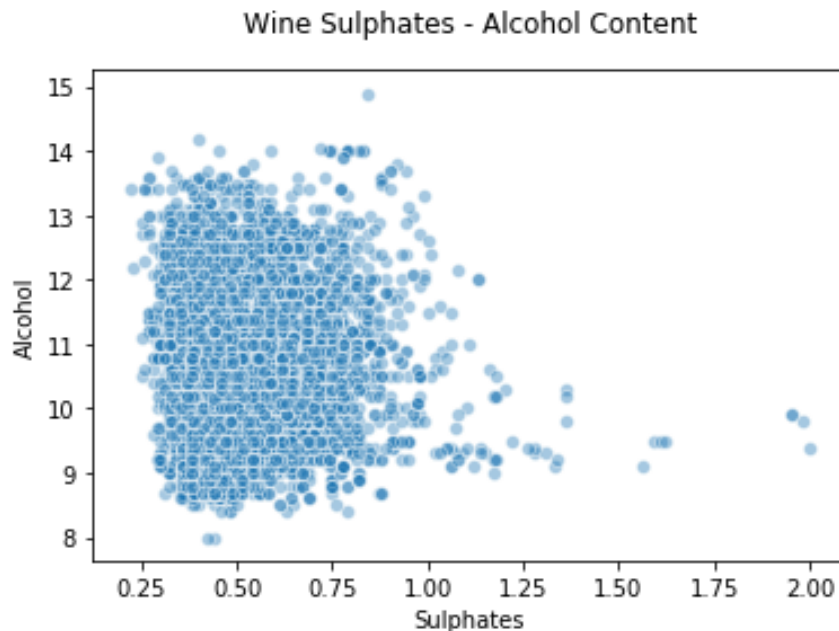
Křivka lokální regrese

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



Kombinovaný graf

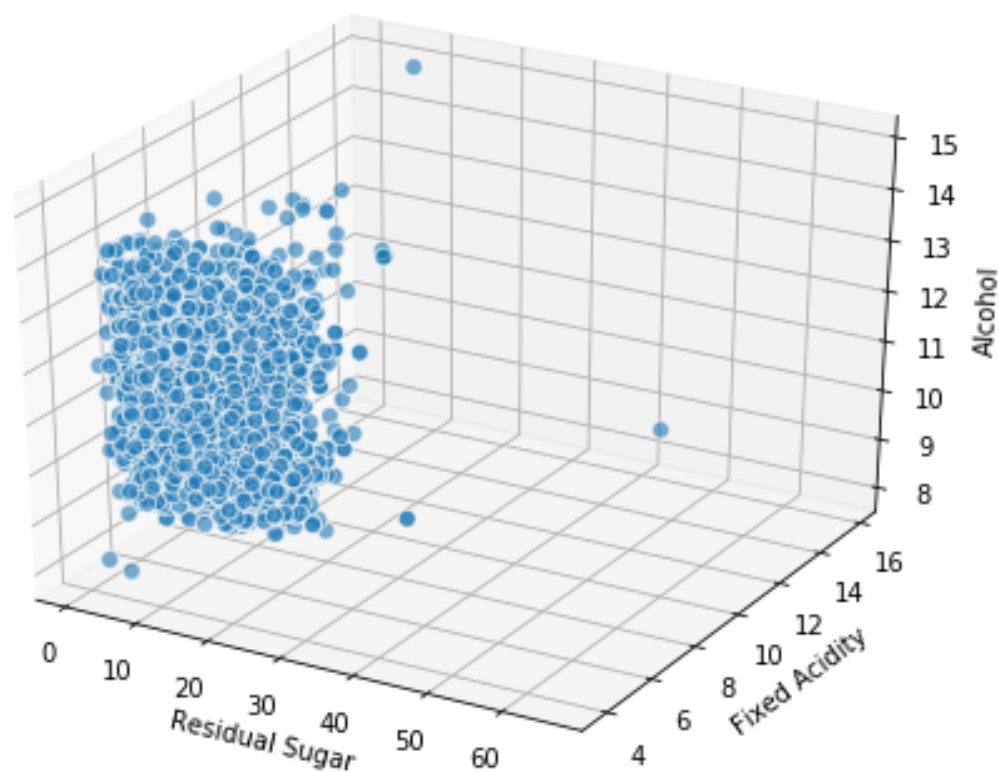
- Kombinuje **bodový graf a graf hustoty** - ukazuje i odhad funkce hustoty pravděpodobnosti



Vizualizační techniky

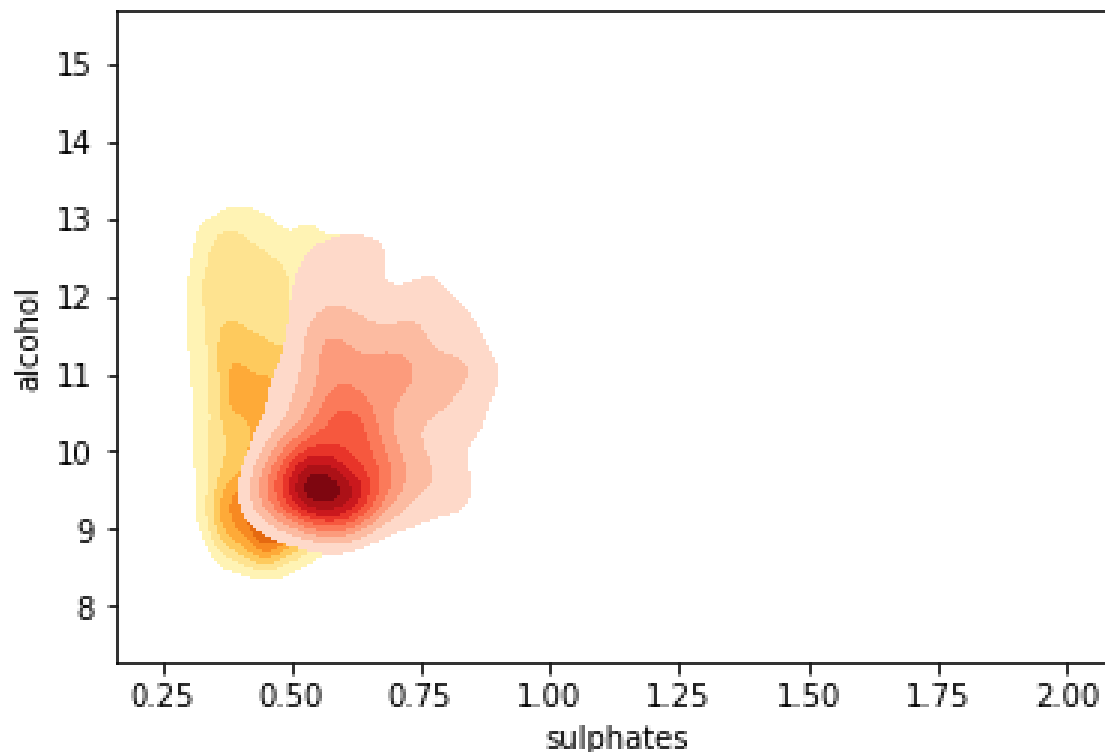
- Vizuální analýza **několika atributů**:
 - Metody **rozšíření** z analýzy jednoho atributu (pokrač.)
 - Vizualizace dalšího atributu použitím nějakého **vizuálního rysu**
 - Hloubky (3D graf)
 - Barevné škály/odstínů
 - Různé velikosti zobrazovaného *bodu*
 - Různého tvaru zobrazovaného *bodu*

Využití hloubky (3D graf)

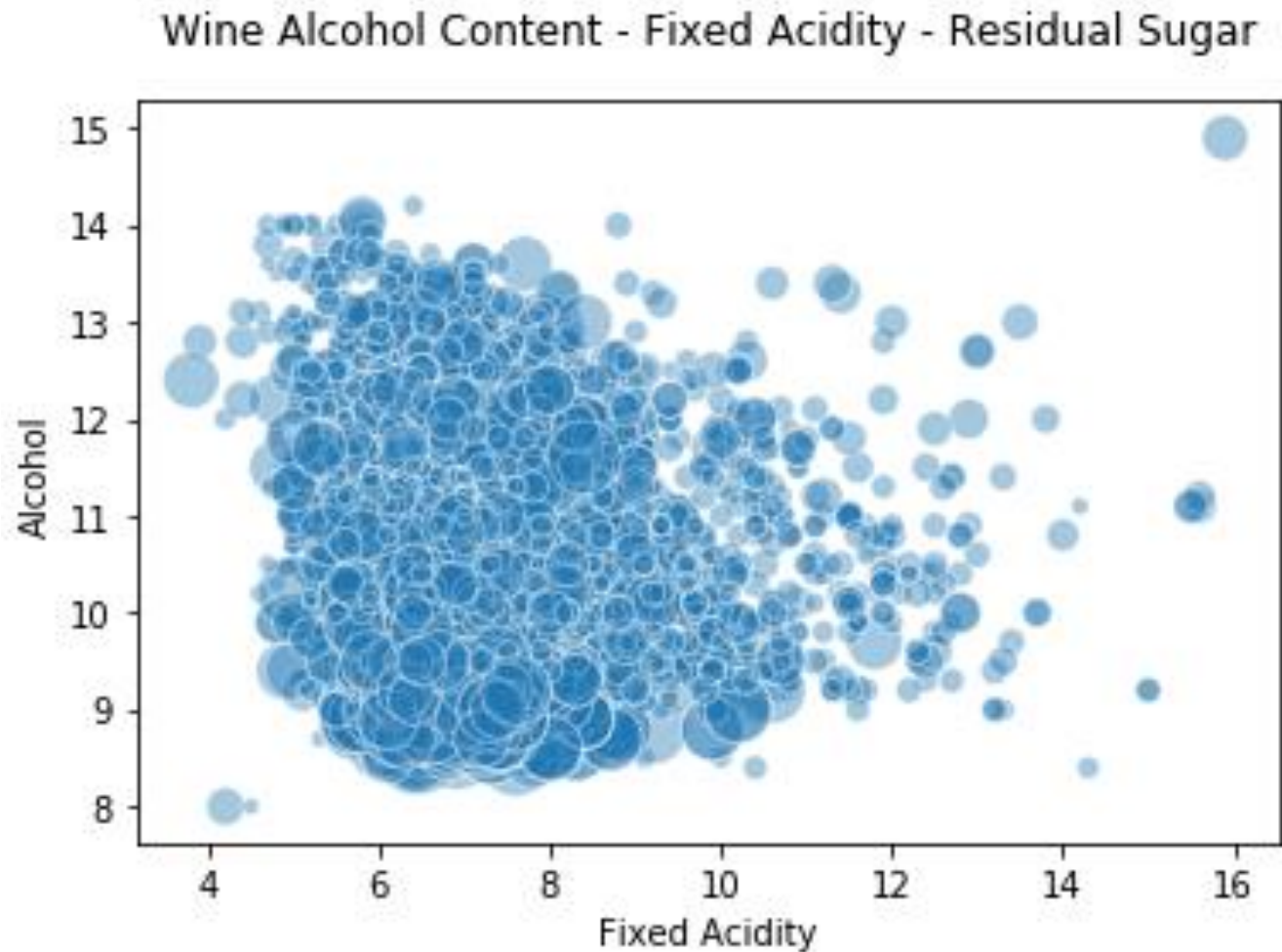


Využití odstínu

- Barva a její odstíny pro kategorický nebo diskretizovaný kvantitativní atribut – zde diskretizovaný odhad funkce hustoty pravděpodobnosti



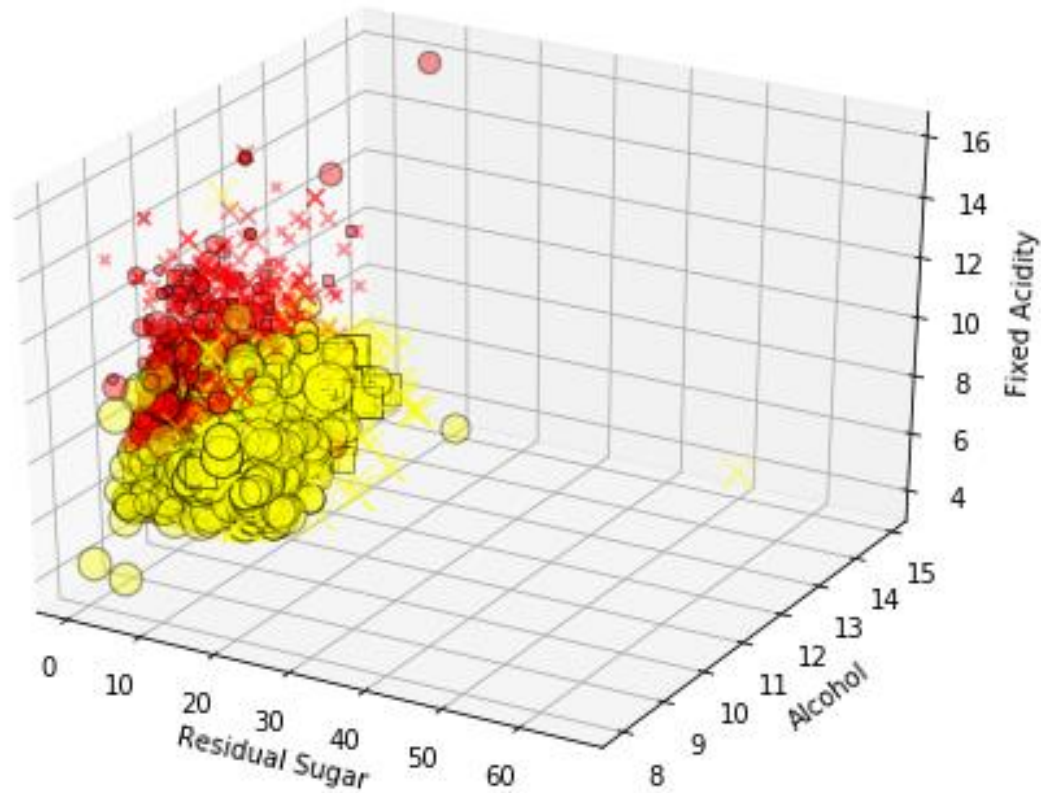
Využití velikosti



Kombinace vizuálních rysů

- Rozšíření využitím **hloubky**, **barvy**, **velikosti** a **tvaru**

Wine Residual Sugar - Alcohol Content - Acidity - Total Sulfur Dioxide - Type - Quality

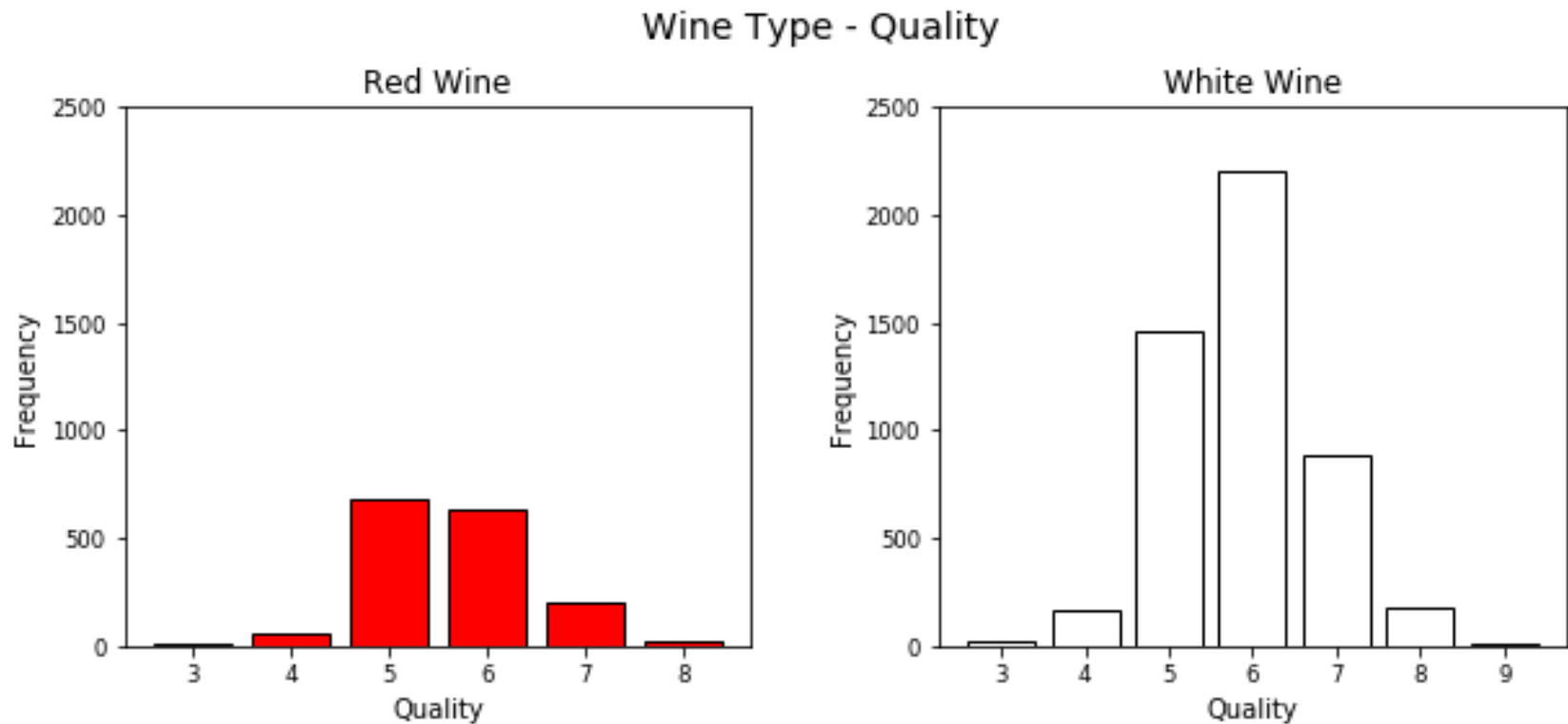


Vizualizační techniky

- Vizuální analýza **několika atributů**:
 - Metody **rozšíření** z analýzy jednoho atributu (pokrač.)
 - Sestavy grafů
 - Oddíly (fasety)/podgrafy
 - Matice grafů
 - Speciální grafy
 - Systém paralelních souřadnic

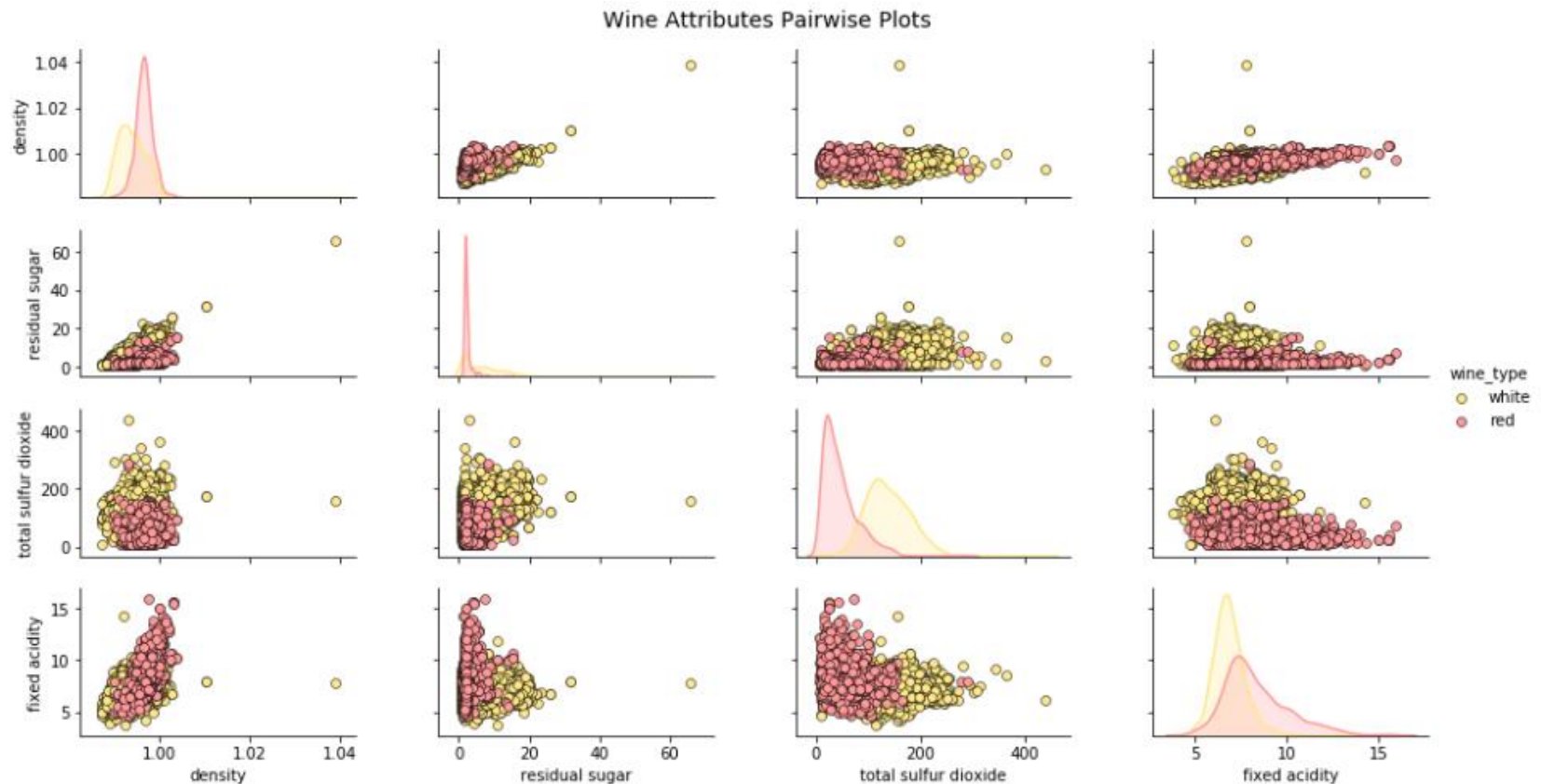
Oddíly (facets)

- Rozčlenění prostoru a umístění podgrafů, zde pro kategorické atributy



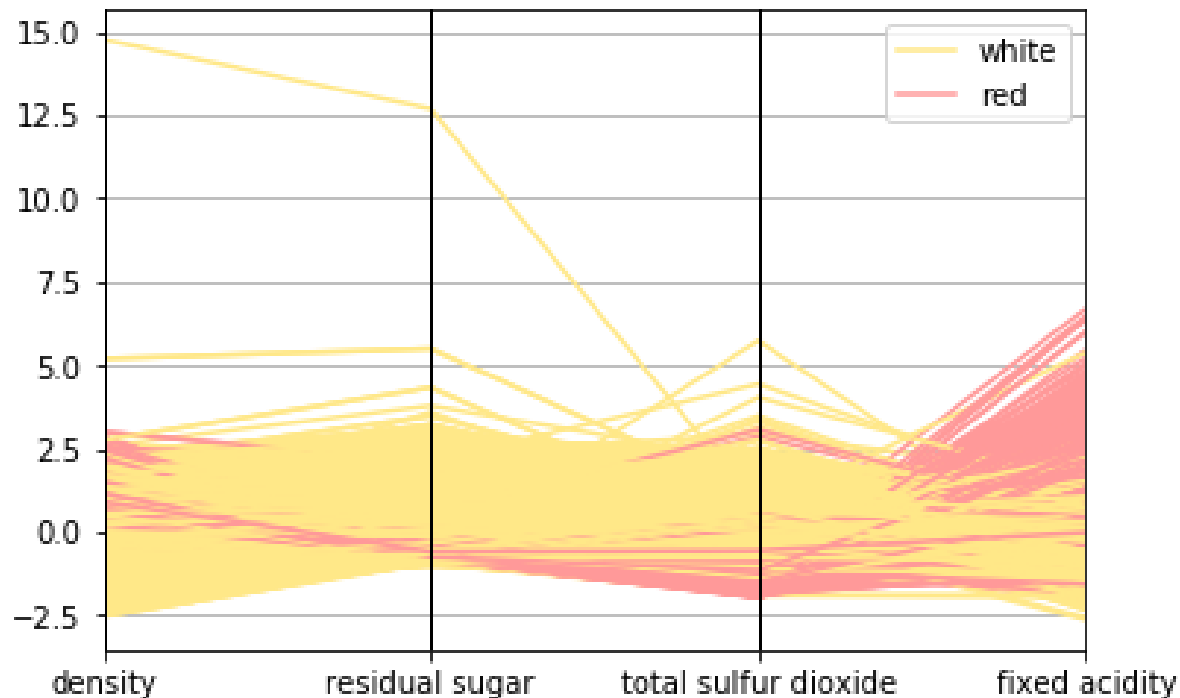
Matice grafů

- Zobrazuje grafy pro všechny dvojice atributů
 - Grafy pod a nad diagonálou symetrické (jen osy přehozené), v diagonále může být odhad funkce hustoty pravděpodobnosti.



System paralelních souřadnic

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



Obsah

- Krok porozumění datům
- Datová sada (dataset)
- Popisné charakteristiky dat
- Vizualizační techniky
- Korelační analýza

Korelační analýza

- Vyjadřuje **statistickou závislost** dvou atributů.
- Přístup
 - Číselné vyjádření
 - Kvantitativní atributy
 - Pearsonův korelační koeficient
 - Spearmanův koeficient korelace pořadí
 - Kategorické atributy
 - Test dobré shody
 - Spearmanův koeficient korelace pořadí (jen pro ordinální)
 - Vizuální techniky
 - Bodový graf
 - Matice korelačních koeficientů s tepelnou mapou
 - Matice bodových grafů

Kvantitativní atributy

□ 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 $\langle -1, 1 \rangle$.

Uspořádatelná data

□ Pro ordinální a kvantitativní atributy

■ Spearmanův koeficient korelace pořadí

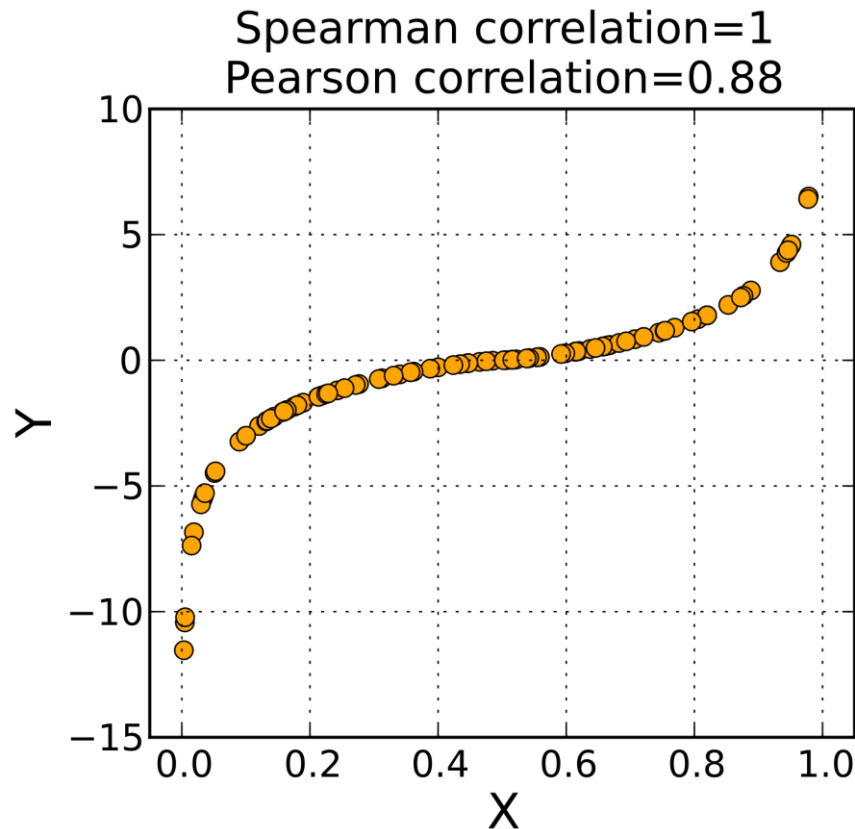
- Založený na pořadí uspořádaných hodnot dvou atributů
- Vyjadřuje sílu a směr **monotónního** vztahu
- **Nejsou-li duplicitní** hodnoty, určí se pro atributy A a B jako

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

- kde n je počet n-tic a $d_i = \text{poradi}(A_i) - \text{poradi}(B_i)$ 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.

Korelační analýza

□ Lineární vs. monotónní korelace



Zdroj: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Kategorická data

- χ^2 - test dobré shody: H_0 – 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.
- Korelace (obecně) neimplikuje kauzalitu
 - *Příklad:* Např. zjištěná korelace atributů '*počet nemocnic*' a '*počet ukradených aut*' ve městě.

Kategorická data – příklad

Určit korelaci binárních atributů *Hraje_šachy* a *Preferuje_sci-fi*.

Kontingenční tabulka:

	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

pozorovaná

očekávaná

Výpočet χ^2 :

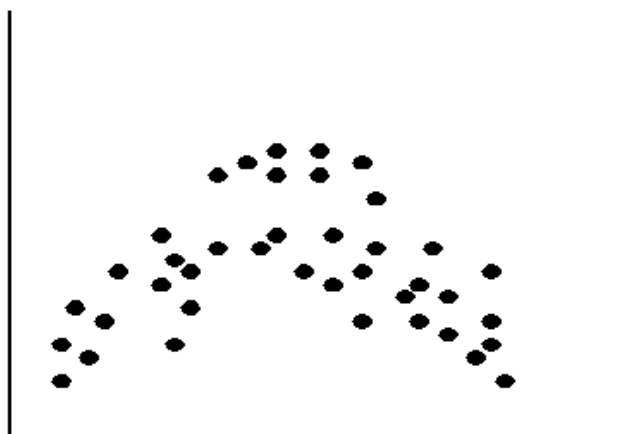
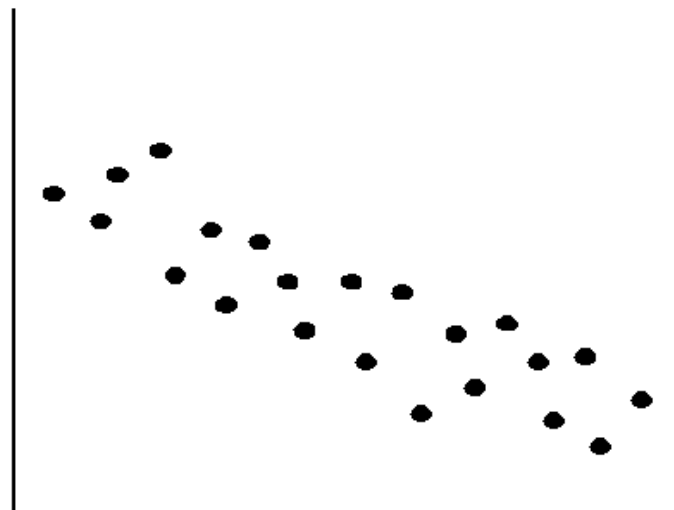
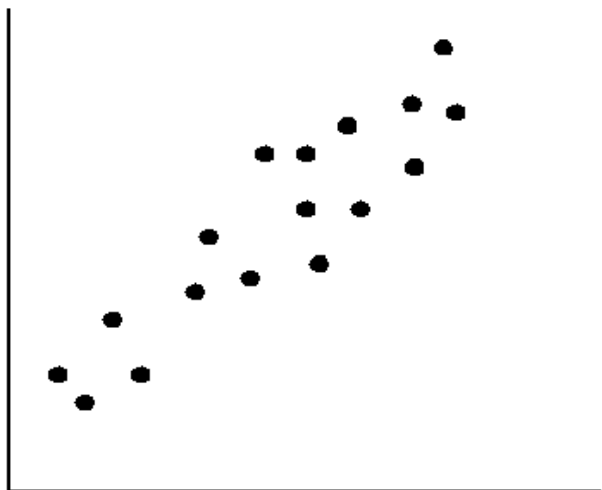
$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93$$

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

→ atributy **jsou korelované**

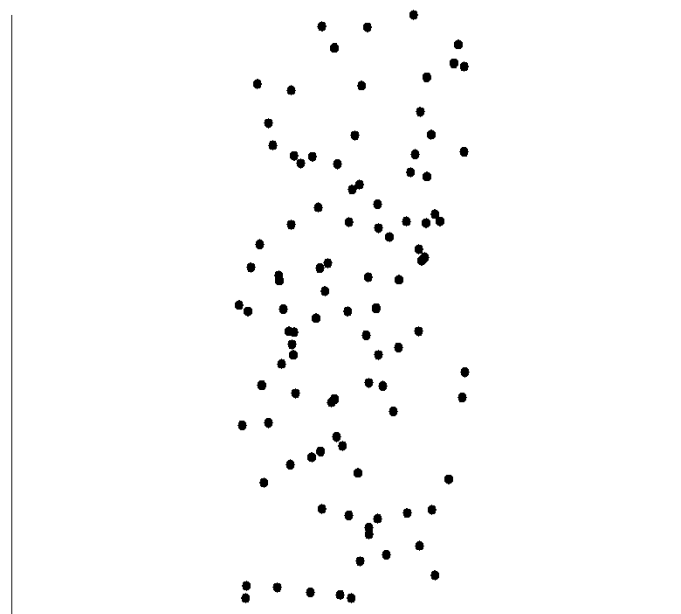
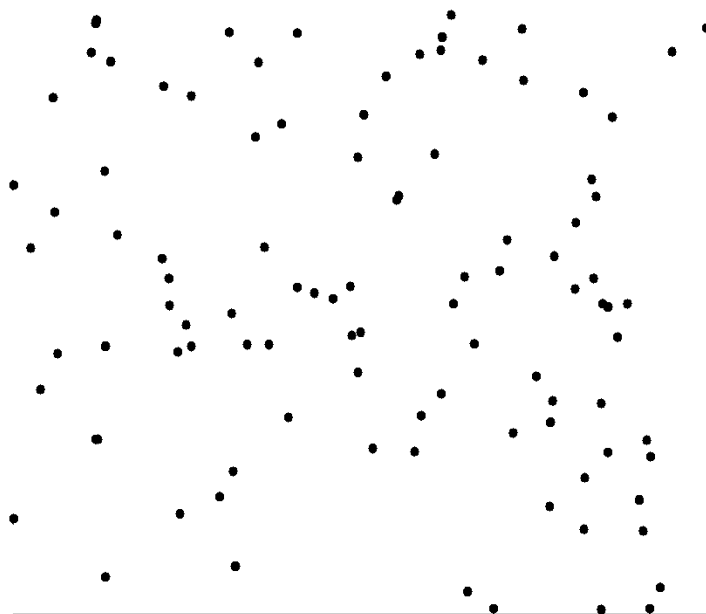
Bodový graf

■ Pozitivní a negativní korelace

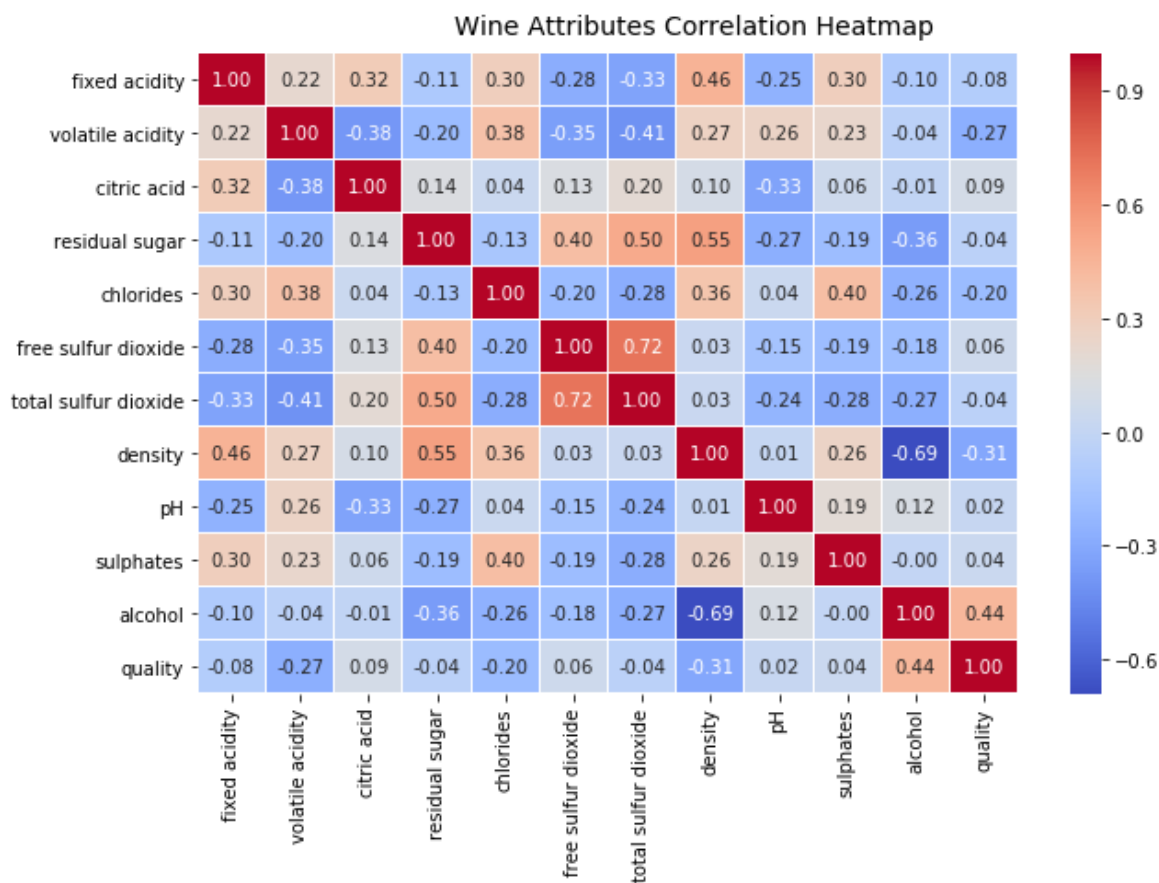


Bodový graf

■ Nekorelovaná data

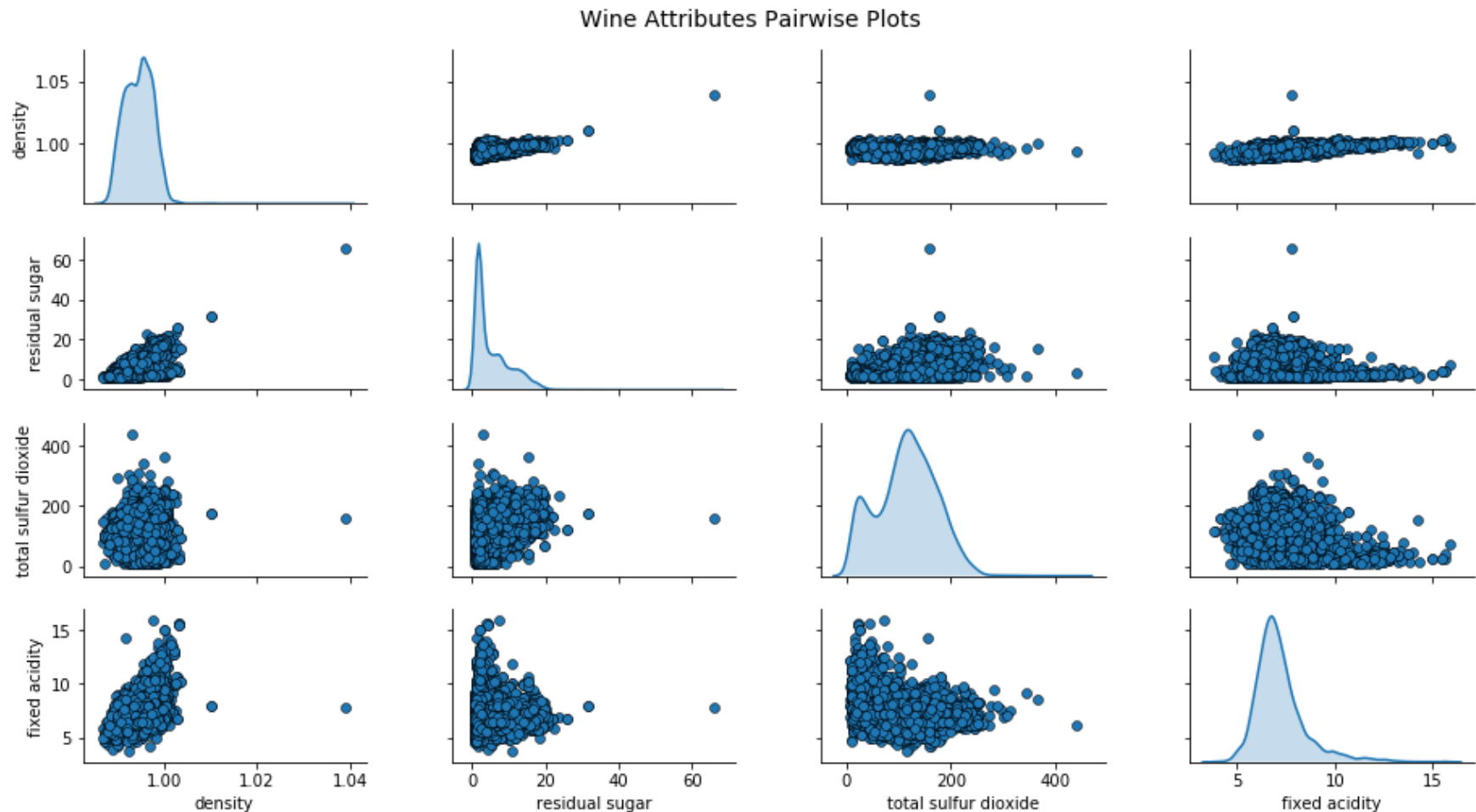


Korelační matice



□ Použití: celkový náhled na vztahy atributů.

Matice bodových grafů



Literatura (1)

1. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Third Edition. Elsevier Inc., 2012, ISBN 978-0-12-381479-1, s. 39 – 81.
2. Zendulka a kol.: Získávání znalostí z databází. Studijní opora, FIT VUT v Brně, 160 s., 2009. (elektronicky) Dostupná mezi soubory k předmětu. (s. 32 – 44).
3. Provost, F., Fawcett, T.: Data Science for Business. O'Reilly Media Inc., 2013. Dostupné na [https://www.researchgate.net/publication/256438799 Data Science for Business](https://www.researchgate.net/publication/256438799_Data_Science_for_Business) .

Literatura (2)

4. Sarkar, D.: The Art of Effective Visualization of Multi-dimensional Data. Dostupné na <https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57>
5. Kinha, Y: Understand the data better by Visualisation. Dostupné na <https://www.edvancer.in/data-visualisation-descriptive-statistics>.