## Latihan/Kuis - 5 menit

- Tentukan apakah aktivitas-aktivitas berikut adalah termasuk dalam lingkup Datamining
  - Membagi pegawai suatu perusahaan berdasarkan gender
  - Memprediksi harga barang pada bulan selanjutnya berdasarkan data historis (bulan-bulan sebelumnya)
  - Memonitor detak jantung pasien untuk mendeteksi kondisi abnormal
  - Menghitung total penjualan suatu perusahaan
  - Mengkarakterisasi pelanggan setia suatu produk



**Computer Science Department**
**Bogor Agricultural University**

"getting familiar with your data"

# Data dan Eksplorasi Data

Kuliah 2

## Outline

- Data dan Tipe data
- Kualitas data
- Statistika ringkasan
- Visualisasi
- Ukuran kemiripan dan ketidakmiripan

- Catatan: semua slide diambil dari
  - Tan P., Michael S., & Vipin K. 2006. *Introduction to Data mining*. Pearson Education, Inc.
  - Han J & Kamber M. 2006. *Data mining – Concept and Techniques.* Morgan-Kauffman, San Diego



## Pertanyaan-pertanyaan yang akan dijawab pada kuliah hari ini

- Apa saja tipe atribut yang menyusun data?
- Apa jenis nilai dari atribut?
- Mana atribut yang diskrit, mana yang kontinu?
- Bagaimana distribusi nilainya?
- Dapatkah kita mendapatkan info tentang pencilan?
- Apakah ada cara memvisualisasikannya?
- Dapatkan kita mengukur similaritas antar objek?

## Jawaban

- Apa saja tipe atribut yang menyusun data?
  - Penjelasan mengenai berbagai tipe atribut
- Apa jenis nilai dari atribut? Mana atribut yang diskrit, mana yang kontinu? Bagaimana distribusi nilainya? Dapatkah kita mendapatkan info tentang pencilan?
  - Penjelasan mengenai dasar-dasar statistika

"Mengetahui dasar-dasar statistika yang sesuai dengan masing-masing atribut akan membuat lebih mudah melakukan **pengisian *missing values***, **menghaluskan nilai-nilai noise**, dan **menunjukkan outliers** selama praproses data"

"Mengetahui atribut dan nilai atributnya akan membantu dalam memperbaiki inkonsistensi yang terjadi selama integrasi data



## Jawaban

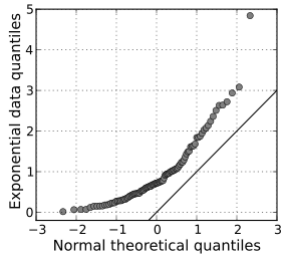
- Apakah ada cara memvisualisasikannya?
  - Distribution plots, quantile plots, histograms, scatter plots, dll

"Memplot kecenderungan (tendency) dari distribusi data akan menunjukkan apakah data tersebut bersifat simetris atau *skewed* (condong/tidak simetris)

"Visualisasi dalam bentuk grafik dapat membantu mengidentifikasi hubungan antar data, tren, dan bias yang tersembunyi dalam dataset yang tidak terstruktur"

## Contoh

- Quantile plots

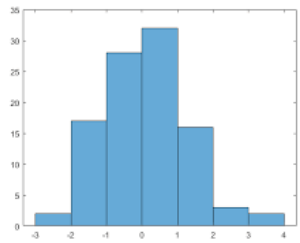


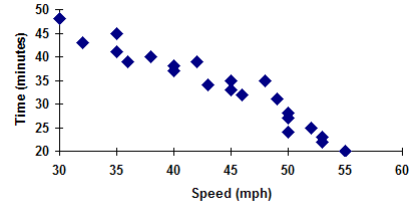

Jika dua dataset memiliki distribusi yang similar, titik-titik pada Q–Q plot akan mendekati garis $y = x$.

Jika distribusinya memiliki hubungan yang linier titik-titik pada Q–Q plot akan terletak pada suatu garis tapi tidak harus pada garis $y = x$.



- Grafik lainnya






Penjelasn rinci di slide-slide selanjutnya

## Jawaban

- Dapatkan kita mengukur similaritas antar objek?
  - Ada banyak ukuran untuk mendapatkan nilai similaritas dan disimilaritas. Pada umumnya, beberapa ukuran merujuk pada ukuran kedekatan
  - Kedekatan dua objek dapat diukur menggunakan fungsi jarak antara atribut-atributnya atau dapat dihitung dengan probabilitas



- **Data dan Tipe data**
- Kualitas data
- Statistika ringkasan
- Visualisasi
- Ukuran kemiripan dan ketidakmiripan

## Apakah Data itu ?

- Koleksi objek data dan atributnya
- Sebuah atribut adalah suatu sifat instriksik atau karakteristik dari suatu objek data
  - contoh: warna mata seseorang, temperatur, dll.
  - Atribut bisa disebut juga sebagai variabel, field, karakteristik, atau fitur
- Kumpulan atribut mendeskripsikan suatu objek
  - Objek dikenal juga sebagai record, titik (point), kasus, sampel, entitas, atau instan

Attributes

Objects

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

## Attribute values (Nilai dari atribut)

- Nilai dari suatu atribut dapat berupa *numbers* (angka) atau simbol (karakter, string) yang diberikan ke suatu atribut
  - Contoh nilai atribut (attribute value):
    - G6543123 (untuk NRP)
    - 36.5 (untuk temperatur)
    - Yes/no (untuk status pernikahan)

## Hal Penting tentang Atribut dan Nilai Atribut

- Atribut yang sama dapat dipetakan pada nilai atribut yang berbeda
  - Contoh: tinggi dapat diukur dalam feet atau meter
- Atribut yang berbeda dapat dipetakan ke dalam himpunan nilai yang sama
  - Contoh: nilai atribut untuk ID dan umur adalah bertipe integer
  - Namun karakteristik nilai atributnya berbeda
    - ID tidak memiliki batasan
    - Umur punya batas minimum dan maksimum



## Pembagian Atribut berdasarkan Tipenya

- Tipe Atribut Katagorikal (Kualitatif)
  - Nominal
  - Ordinal
- Tipe Atribut Numerik
  - Interval
  - Rasio

## Tipe Atribut : Katagorikal (Kualitatif)

| Tipe Atribut | Deskripsi | Contoh | Operasi |
|---|---|---|---|
| Nominal | Nilai dari atribut nominal hanya berupa nama-nama yang berbeda.<br><br>Atribut nominal hanya menyediakan informasi yang cukup untuk membedakan suatu objek dengan objek yang lain (=, ≠) | Kode pos, ID pegawai, warna mata, gender: {*pria, wanita*} | modus, entropy, contingency correlation, $\chi^2$ test |
| Ordinal | Nilai atribut ordinal menyediakan informasi yang cukup untuk MENGURUTKAN objek (<, >) | Tingkat kekerasan material {*good, better, best*}, rangking, nomor jalan | median, percentiles, rank correlation, run tests, sign tests |

## Tipe Atribut: Numerik (Kuantitatif)

| Tipe Atribut | Deskripsi | Contoh | Operasi |
|---|---|---|---|
| Interval | Pada tipe atribut interval perbedaan antar nilai-nilainya memiliki makna. Untuk pengukuran, misal: celcius atau fahrenheit, (+, - ) | dates (penanggalan), temperatur dalam Celsius atau Fahrenheit | mean, standard deviation, Pearson's correlation, *t* and *F* tests |
| Ratio | Untuk tipe atribut rasio, Perbedaan nilai dan rasio dari nilai-nilai atributnya memiliki makna (*, /) | Temperatur dalam Kelvin, monetary quantities, counts, age, mass, panjang, arus listrik | geometric mean, harmonic mean, percent variation |

## Contoh

- Tipe atribut interval
  - Perbedaan panas antara temperatur pada 90° C dan 100° C sama dengan perbedaan temperatur 80° C dan 90° C
  - Tapi panas pada temperatur 100° C tidak berarti dua kali dari panas temperatur 50° C
  - 0° C tidak berarti tidak ada panas
- Tipe atribut rasio
  - panas pada temperatur 100° K berarti dua kali dari panas temperatur 50° K
  - 0° K berarti memang tidak ada panas



## Untuk atribut di bawah ini, tipe atributnya apa?

- PH (untuk mengukur tingkat keasaman)
- Berat badan
- Fenotipe (warna mata, warna kulit, dll)
- Tingkat rasa sakit
- IPK
- Nilai mata kuliah
- Musim

## Pembagian Atribut berdasarkan Kataktistik Datanya

- Atribut Diskrit
- Atribut Kontinu



## Atribut diskrit dan kontinu

- Atribut Diskrit
  - Hanya memiliki himpunan nilai yang terbatas
  - Contoh: kode pos, counts, atau sekumpulan kata pada suatu koleksi dokumen
  - Biasanya direpresentasikan sebagai variabel integer
  - Catatan: atribut biner (binary attributes) adalah kasus khusus dari atribut diskrit
- Atribut Kontinu
  - Nilai atributnya berupa bilangan riil
  - Contoh: temperatur, tinggi, atau berat.
  - Pada praktiknya, bilangan riil hanya dapat diukur dan direpresentaikan menggunakan sejumlah terbatas digit
  - Biasanya direpresentasikan sebagai variabel float (floating-point variables).

## Termasuk apakah atribut di bawah ini? Atribut Diskrit atau Atribut Kontinu?

- PH (untuk mengukur tingkat keasaman)
- Berat badan
- Fenotipe (warna mata, warna kulit, dll)
- Tingkat rasa sakit
- IPK
- Nilai mata kuliah
- Musim



## Tipe-tipe Dataset

- **Record**
  - **Data Matriks**
  - **Data Dokumen**
  - **Data Transaksi**
- **Graph**
  - **World Wide Web, Twitter, dll**
  - **Stuktut Molekul, Interaksi-interaksi Protein**
- **Ordered (yang bisa diurutkan)**
  - **Data spasial**
  - **Data temporal**
  - **Data Sekuensial**
  - **Data sekuens DNA/genetik**

## Record Data

- Data yang terdiri atas kumpulan record, di mana masing-masing memiliki himpunan atribut

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

22

## Data Transaksi

- Tipe khusus dari record data, di mana
  - Tiap record atau (transaksi) melibatkan sekumpulan item
  - Contoh, toko grosir berikut. Tabel di bawah menunjukkan satu set produk yang dibeli oleh Pembeli dalam satu kali pembelian.

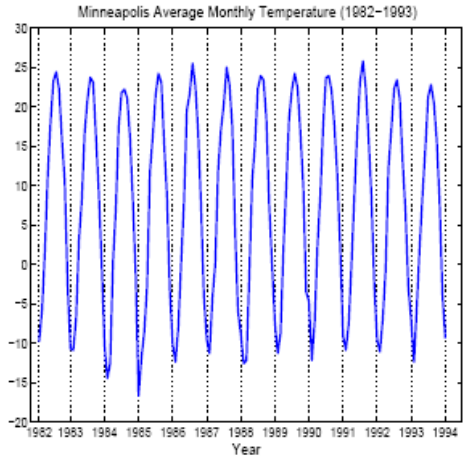
    Satu set pembelian tersebut dinamakan TRANSAKSI, adapun masing-masing produk disebut item

| *TID* | *Items* |
|---|---|
| 1 | Bread, Tea, Milk |
| 2 | Coffee, Bread |
| 3 | Coffee, Tea, Sugar, Milk |
| 4 | Coffee, Bread, Sugar, Milk |
| 5 | Tea, Sugar, Milk |

23

## Ordered Data (Data yang dapat diurutkan)

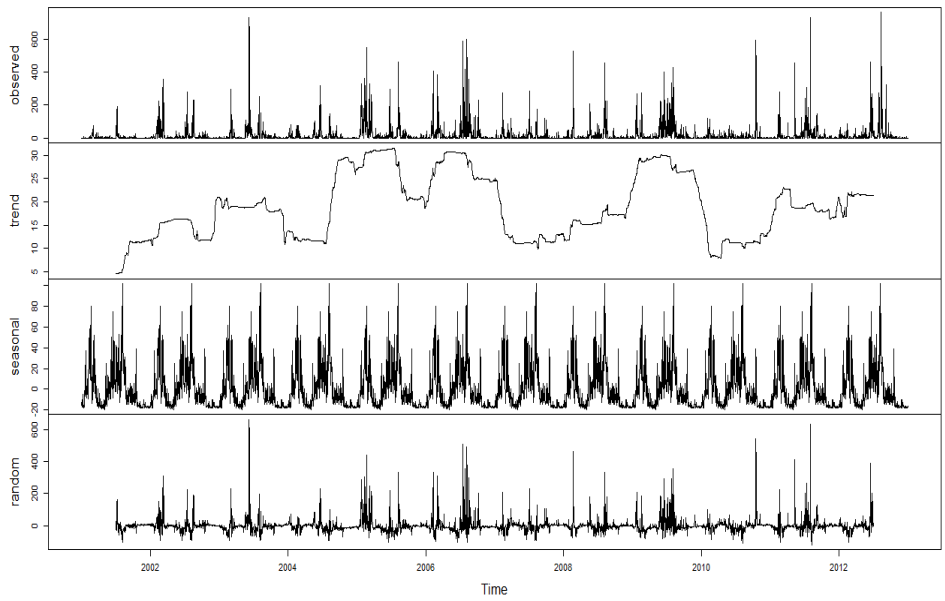
Data temporal




- Data temporal adalah data yang objeknya memiliki atribut yang mereprensentasikan pengukuran yang dilakukan dari waktu ke waktu
- Contoh, data kebakaran hutan yang memberikan informasi terjadinya perambatan kebakaran dari hari ke hari
- Suatu time series adalah suatu sekuen (barisan) pengukuran dari beberapa atribut

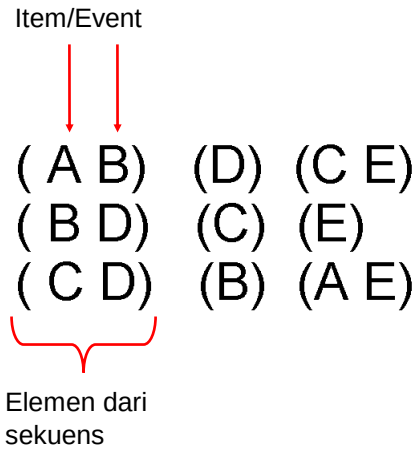
24

## Time series data - hotspot data



## Ordered Data

- Sekuen Transaksi




Data masih mengandung suatu set transaksi dan item, namun atribut waktu dan ID pembeli telah diasosiasikan terhadap masing-masing transaksi



- Data dan Tipe data
- **Kualitas data**
- Statistika ringkasan
- Visualisasi
- Ukuran kemiripan dan ketidakmiripan

## Kualitas Data

Isu yang penting terkait Kualitas Data:

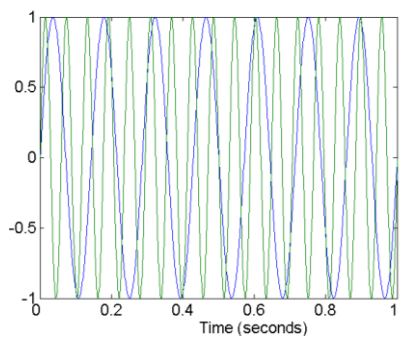
- Problem apa yang muncul terkait dengan kualitas data?
- Bagaimana kita mendeteksi problem tersebut?
- Apa yang bisa kita lakukan untuk mengatasinya?

- Contoh problem yang terkait dengan Kualitas Data:
  - Gangguan (Noise) dan pencilan (outliers)
  - Nilai yang hilang (missing values)
  - Duplikasi data

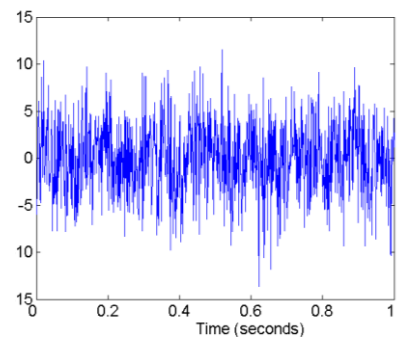
28

## Noise

- Noise terkait dengan berubahnya nilai atribut yang asli (original)
- Contoh: distorsi suara manusia ketika berbicara di tengah keramaian, atau pada telepon yang jelek, atau ada suara berisik dari televisi




Two Sine Waves

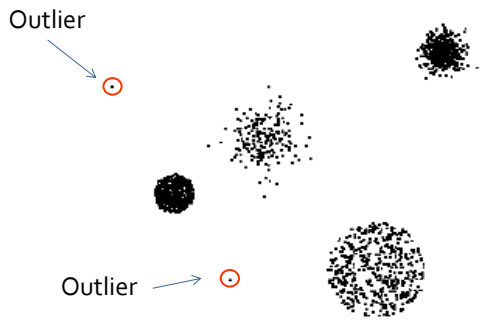



Two Sine Waves + Noise

29

## Pencilan (Outliers)

- Outliers adalah objek data yang memiliki karakteristik yang sangat berbeda dengan objek data pada umumnya pada suatu dataset




Dari pertemuan 1

Outlier: objek data yang **tidak mengikuti perilaku umum** dari data

Dapat dipandang sebagai ***noise*** atau eksepsi tetapi berguna dalam *fraud detection*, *rare events analysis*

30

## Missing Values

- Sebab-sebab missing values
  - Informasi tersebut tidak diperoleh (tidak terkumpul)
    (contoh, orang menolak memberikan info umur dan berat badan)
  - Atribut tidak dapat diterapkan pada semua kasus
    (e.g., atribut pendapatan tahunan tidak bisa dijawab/tidak cocok dengan anak-anak)

- Penanganan missing values
  - Mengeliminasi Objek Data
  - Mengestimasi Missing Values, dengan melakukan interpolasi menggunakan nilai-nilai yang ada pada atribut tersebut
  - Mengabaikan Missing Value selama tahap analisis
    - misal, objek-2 data akan diklusterkan dan perlu menghitung similaritas antar pasangan objek data
    - Similaritas ini dapat dihitung dengan hanya menggunakan nilai data yang ada

31

## Duplikasi Data

- Dataset mungkin mengandung objek-2 data yang sama (duplikat) atau hampir serupa
  - Terjadi jika melakukan merge (penggabungan) dari berbagai sumber data
  - Contoh: orang yang sama dengan banyak alamat email
- Perhatian perlu diberikan untuk menhindari mengkombinasikan objek yang similar tapi tidak duplikat
- Data cleaning (Pembersihan Data)
  - Proses yang terkait dengan isu duplikasi data
  - Menghilangkan *noise* dan data yang tidak konsisten



- Data dan Tipe data
- Kualitas data
- **Statistika ringkasan**
- Visualisasi
- Ukuran kemiripan dan ketidakmiripan

## Summary Statistics (Statistika Ringkasan)

- Statistika ringkasan adalah angka yang men-summary-kan (meringkas) properti (karakteristik) data
  - Peringkasan properti/karakteristik (Summarized properties) meliputi frekuensi, lokasi (location), dan penyebaran (spread)
    - Examples: lokasi – mean, median
      penyebaran - standard deviation
  - Sebagian besar statistika ringkasan dapat dihitung langsung dari data

## Frequency and Mode (Frekuensi dan Modus)

- Frekuensi dari suatu nilai atribut adalah persentase kemunculan nilai tersebut di dalam dataset
  - Contoh, diberikan atribut 'gender' dan suatu representasi populasi dari manusia, kemunculan gender 'female' adalah 50% selama periode waktu tersebut
- Modus adalah nilai atribut yang memiliki frekuensi paling tinggi (yang paling sering muncul)
- Frekuensi biasanya digunakan pada data katagorikal

## Ukuran lokasi: Mean and Median (Rata-rata dan Nilai tengah)

- Mean adalah ukuran yang paling umum dari lokasi suatu himpunan titik
- Namun, mean sangat sensitif terhadap outlier
- Maka, median atau *a trimmed mean* juga umum digunakan

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

## Measures of Spread (Ukuran penyebaran): Range and Variance (Jangkauan dan Ragam)

- Jangkauan adalah perbedaan antara nilai maksimum dan minimum dari data
- Ragam (variance) atau simpangan baku (standard deviation) adalah ukuran paling umum untuk penyebaran dari suatu himpunan titik

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- Namun, ragam juga sensitif terhadap outliers, maka ukuran yang sering digunakan adalah:

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

$$\text{MAD}(x) = \textit{median}\left(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\right)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

AAD: absolute average deviation
MAD: median absolute deviation
IQR: interquartile rage

- Data dan Tipe data
- Kualitas data
- Statistika ringkasan
- **Visualisasi**
- Ukuran kemiripan dan ketidakmiripan



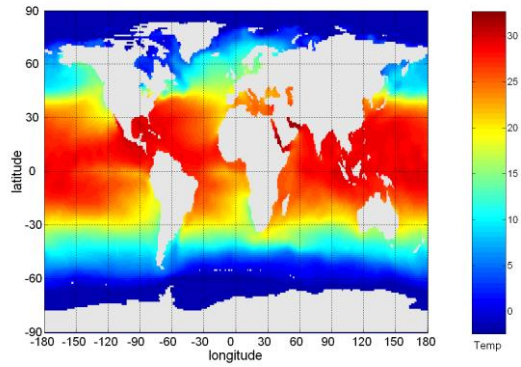
## Visualisasi

Visualisasi adalah konversi data ke dalam format visual atau tabular sehingga karakteristik data dan keterhubungan antar item data atau atribut dapat dianalisis

- Visualisasi data adalah salah satu teknik yang *powerful* dan menarik untuk melakukan eksplorasi data
  - Dapat mendeteksi pola umum dan tren
  - Dapat mendeteksi outlier dan pola yang tidak umum

## Contoh: Temperatur permukaan laut

- Gambar berikut memperlihatkan temperatur permukaan laut (Sea Surface Temperature/SST) bulan Juli 1982
  - Puluhan ribu titik data divisualisasikan dalam satu gambar




## Representasi

- Representasi adalah pemetaan informasi ke dalam format visual
- Objek data, atribut-atributnya, dan keterhubungan antar objek data diterjemahkan ke dalam bentuk elemen grafik, seperti titik, garis, bentuk (shapes) dan warna
- Contoh:
  - Objek data sering direpresentasikan sebagai titik
  - Atribut-atributnya dapat direpresentasikan sebagai posisi dari titik atau karakteristik dari titik, seperti: warna, ukuran, dan bentuk
  - Jika representasi posisi yang digunakan, maka keterhubungan antar titik, misal apakah titik-titik tersebut membentuk kelompok (cluser) atau membentuk outlier, menjadi mudah untuk dilihat

## Arrangement (Pengaturan)

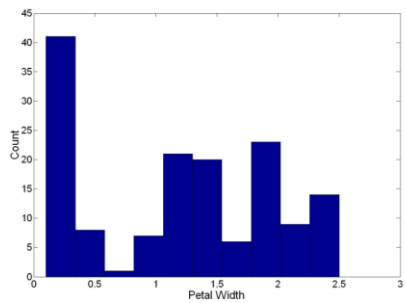
- Pengaturan adalah penempatan elemen-elemen visual dalam suatu tampilan (display)
- Dapat membuat perbedaan yang signifikan dalam bagaimana membaca/memahami data
- Contoh:

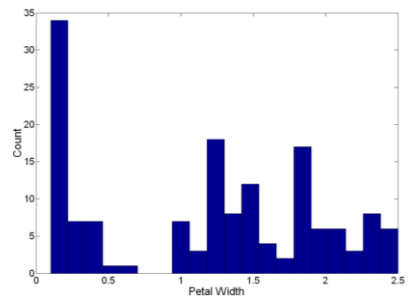
| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| 2 | 1 | 0 | 1 | 0 | 0 | 1 |
| 3 | 0 | 1 | 0 | 1 | 1 | 0 |
| 4 | 1 | 0 | 1 | 0 | 0 | 1 |
| 5 | 0 | 1 | 0 | 1 | 1 | 0 |
| 6 | 1 | 0 | 1 | 0 | 0 | 1 |
| 7 | 0 | 1 | 0 | 1 | 1 | 0 |
| 8 | 1 | 0 | 1 | 0 | 0 | 1 |
| 9 | 0 | 1 | 0 | 1 | 1 | 0 |

| | 6 | 1 | 3 | 2 | 5 | 4 |
|---|---|---|---|---|---|---|
| 4 | 1 | 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 |
| 6 | 1 | 1 | 1 | 0 | 0 | 0 |
| 8 | 1 | 1 | 1 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1 | 1 | 1 |
| 3 | 0 | 0 | 0 | 1 | 1 | 1 |
| 9 | 0 | 0 | 0 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| 7 | 0 | 0 | 0 | 1 | 1 | 1 |

## Teknik visualisasi: Histograms

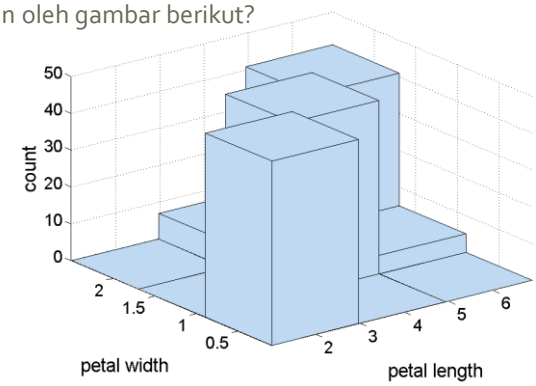
- Histogram
  - Biasanya memperlihatkan distribusi nilai dari suatu variabel
  - Membagi nilai-nilai ke dalam bin dan memperlihatkan plot dalam bentuk bar dari sejumlah objek pada masing-masing bin
  - Tinggi bar menunjukkan jumlah objek
  - Bentuk histogram tergantung dari jumlah bin
- Contoh: lebar Petal (pada kelopak bunga) (10 dan 20 bins)



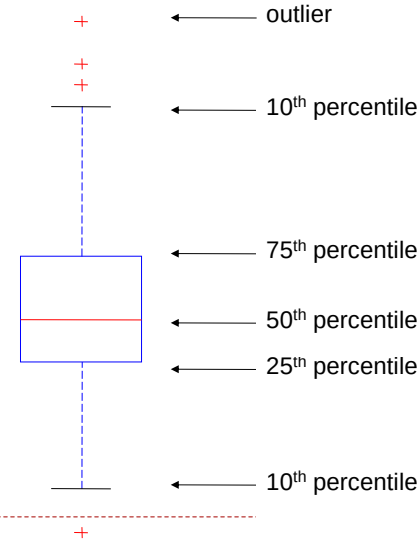
## 2-DHistograms

- Memperlihatkan *joint distribution* dari nilai-nilai dua atribut
- Contoh: lebar petal dan panjang petal
  - Apa yang ditunjukkan oleh gambar berikut?



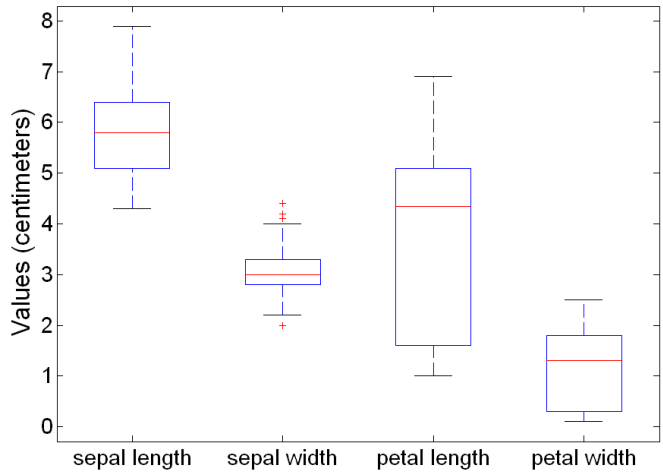

## Teknik visualisasi: Box Plots

- Box Plots
  - Ditemukan olehJ. Tukey
  - Cara lain untuk menampilkan distribusi data
  - Gambar berikut menunjukkan bagian-bagian dasar suatu box plot



## Contoh Box Plots

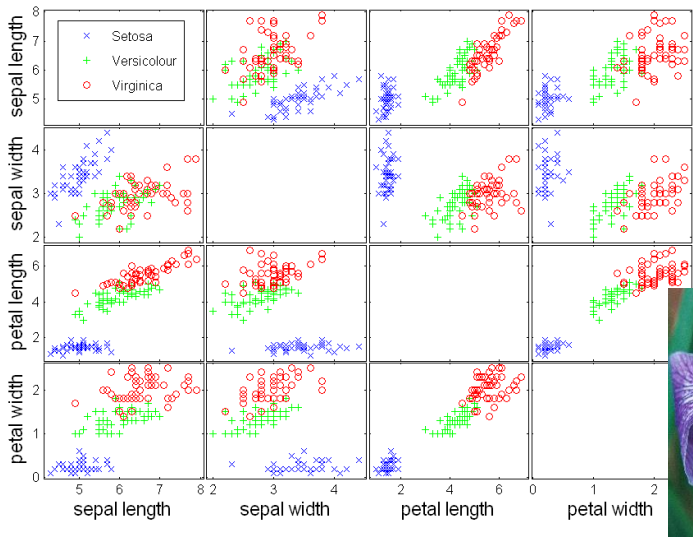
- Box plots dapat digunakan untuk membandingkan atribut-atribut

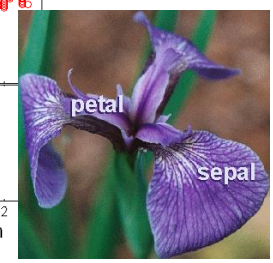



## Teknik Visualisasi: Scatter Plots

- Scatter plots
  - Nilai atribut-atribut menentukan posisi
  - Umumnya 2-D scatter plots, tapi dapat juga digambarkan dalam 3-D scatter plots
  - Sering kali atribut-atribut tambahan dapat ditampilkan menggunakan ukuran, bentuk, dan warna dari penanda (marker) yang merepresentasikan objek
  - Penting untuk memiliki array dari scatter plots yang dapat meringkas beberapa pasangan atribut
    - Lihat contoh di slide selanjutnya

## Scatter Plot Array of Iris Attributes




Source: http://mirlab.org

- Data dan Tipe data
- Kualitas data
- Statistika ringkasan
- Visualisasi
- **Ukuran kemiripan dan ketidakmiripan**

## Kemiripan (Similarity) dan Ketidakmiripan (Dissimilarity)

- Kemiripan (Similarity)
  - Ukuran numerik dari seberapa mirip dua buah objek data
  - Semakin tinggi kemiripan antar dua objek, semakin tinggi nilai kemiripannya
  - Nilai kemiripan biasanya berada pada range [0,1]
- Ketidakmiripan (Dissimilarity)
  - Ukuran numerik yang menggambarkan seberapa berbeda dua buah objek data
  - Semakin tinggi kemiripan antar dua objek, semakin rendah nilai ketidakmiripannya (disimilaritasny)
  - Nilai minimum dari disimilaritas adalah 0
  - Batas atasnya bervariasiUpper limit varies
- Kedekatan (Proximity) mengacu pada kemiripan atau ketidakmiripan

## Kemiripan/Ketidakmiripan untuk atribut sederhana

*p* dan *q* nilai atribut untuk dua buah objek data

| Attribute Type | Dissimilarity | Similarity |
|---|---|---|
| Nominal | $d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$ | $s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$ |
| Ordinal | $d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where $n$ is the number of values) | $s = 1 - \frac{\lvert p-q \rvert}{n-1}$ |
| Interval or Ratio | $d = \lvert p - q \rvert$ | $s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$ |

**Table 5.1.** Similarity and dissimilarity for simple attributes

## Euclidean Distance

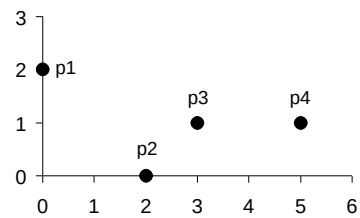
- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

di mana $n$ jumlah dimensi (atribut) dan $p_k$ dan $q_k$ berturut-turut adalah nilai atribut ke-k dari objek data $p$ dan $q$.

- Jika skalanya berbeda, standarisasi diperlukan

## Euclidean Distance




| point | x | y |
|---|---|---|
| p1 | 0 | 2 |
| p2 | 2 | 0 |
| p3 | 3 | 1 |
| p4 | 5 | 1 |

| | p1 | p2 | p3 | p4 |
|---|---|---|---|---|
| p1 | 0 | 2.828 | 3.162 | 5.099 |
| p2 | 2.828 | 0 | 1.414 | 3.162 |
| p3 | 3.162 | 1.414 | 0 | 2 |
| p4 | 5.099 | 3.162 | 2 | 0 |

Distance Matrix

## Minkowski Distance

- Minkowski Distance adalah bentuk umum dari Euclidean Distance

$$dist = \left( \sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Di mana *r* adalah suatu parameter, *n* adalah jumlah dimensi (attributes); $p_k$ dan $q_k$ adalah atribut ke-k dari objek data *p* dan *q*

## Mahalanobis Distance

$$mahalanobis(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$

Σ adalah matriks kovarian data masukan *X*

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

## Karakteristik umum jarak (Distance)

- Jarak, seperti misalnya Euclidean distance, memiliki karakteristik berikut:
  1. $d(p, q) \geq 0$ untuk semua *p* dan *q* dan $d(p, q) = 0$ hanya jika $p = q$ → Positive definiteness
  2. $d(p, q) = d(q, p)$ untuk semua *p* dan *q* → Symmetri
  3. $d(p, r) \leq d(p, q) + d(q, r)$ untuk semua titik *p*, *q*, dan *r* → Triangle Inequality

  di mana $d(p, q)$ adalah jarak (ketidakmiripan/dissimilarity) antara titik-titik (objek data), *p* dan *q*.

- Suatu jarak yang memenuhi karakteristik di atas adalah suatu metrik (metric)

## Karakteristik Umum Kemiripan (Similarity)

- Kemiripan (Similarities), memiliki beberapa karakteristik:
  1. *s(p, q) = 1* atau kemiripan maksimal (maximum similarity) hanya jika *p = q*.
  2. *s(p, q) = s(q, p)* untuk semua *p* dan *q* → Simetri

  di mana *s(p, q)* adalah similaritas antara titik-titik (objek data) *p* dan *q*.

## Kemiripan (Similarity) antara vektor biner (Binary Vectors)

- Situasi yang umum adalah objek *p* dan *q* hanya memiliki atribut biner
- Hitung kemiripan dengan menggunakan ketentuan berikut:
  $M_{01}$ = jumlah atribut di mana p bernilai 0 dan q bernilai 1
  $M_{10}$ = jumlah atribut di mana p bernilai 1 dan q bernilai 0
  $M_{00}$ = jumlah atribut di mana p bernilai 0 dan q bernilai 0
  $M_{11}$ = jumlah atribut di mana p bernilai 1 dan q bernilai 1
- Simple Matching dan Koefisien Jaccard
  SMC = jumlah yang matches / jumlah atribut
  = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = jumlah atribut 11 matches (both matches)/jumlah nilai atribut yang not-both-zero
  = $(M_{11}) / (M_{01} + M_{10} + M_{11})$

## SMC versus Jaccard: Contoh

*p* = 1 0 0 0 0 0 0 0 0 0

*q* = 0 0 0 0 0 0 1 0 0 1

$M_{01}$ = 2 (jumlah atribut di mana p bernilai 0 dan q bernilai 1)
$M_{10}$ = 1 (jumlah atribut di mana p bernilai 1 dan q bernilai 0)
$M_{00}$ = 7 (jumlah atribut di mana p bernilai 0 dan q bernilai 0)
$M_{11}$ = 0 (jumlah atribut di mana p bernilai 1 dan q bernilai 1)

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (0+7) / (2+1+0+7) = 0.7

J = $(M_{11})$ / $(M_{01} + M_{10} + M_{11})$ = 0 / (2 + 1 + 0) = 0

## Ringkasan Materi

- Data dan Tipe data
- Kualitas data

- Statistika ringkasan
- Visualisasi
- Ukuran kemiripan dan ketidakmiripan

## Ringkasan Materi hari ini !

- Data dan Tipe data
  - Definisi : data, objek data, atribut, nilai atribut
  - Pembagian Atribut Berdasarkan Tipenya : katagorikal (kualitatif) – nominal ordinal, numerik (kuantitatif) – interval, rasio
  - Pembagian Atribut Berdasarkan Karakteristik Datanya: diskrit, kontinu
  - Tipe-tipe Dataset: record, graph, ordered data
- Kualitas data
  - Noise
  - Outlier
  - Duplikasi Data
  - Missing Value
- Statistika ringkasan
  - Frekuensi – model
  - Lokasi – mean, median
  - Sebaran – Standard deviasi
- Visualisasi
- Ukuran kemiripan dan ketidakmiripan



## Exercises

- Classify the following attributes as binary, discrete, or continuous, qualitative (nominal or ordinal) or quantitative (interval or ratio)

| Attribute | Binary | Discrete | Continuous | qualitative | quantitative |
|---|---|---|---|---|---|
| Distance to nearest hospital | | | | | |
| Temperature of human body | | | | | |
| Temperature of human body By human judgments | | | | | |
| Exam grade | | | | | |
| Predicate of graduation | | | | | |

Materi selanjutnya:

# Praproses data