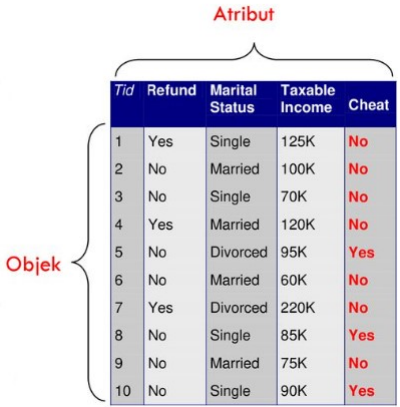
Exploratory Data Anaysis
Session 1

SSD23407

Egi Safitri, S.Mat., M.Si

Apa itu Set Data?

1. Set data adalah kumpulan dari objek data dan atributnya.
2. Sebuah atribut adalah sifat atau karakteristik dari sebuah objek.
   - Contoh: warna mata dari seseorang, temperatur suhu
   - Atribut juga dikenal sebagai variabel, karakteristik, atau fitur
3. Koleksi dari atribut mendeskripsikan sebuah objek. Objek juga dikenal sebagai record, point, case, sample, entity, atau instance.



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

Nilai Atribut

1. Nilai atribut adalah angka atau simbol yang ditetapkan untuk sebuah atribut.
2. Perbedaan antara atribut dan nilai atribut
   - Atribut yang sama bisa dipetakan ke nilai atribut yang berbeda. Contoh : Tinggi badan dapat dihitung dalam meter atau feet
   - Atribut yang berbeda dapat dipetakan ke nilai atribut yang sama. Contoh : Nilai atribut untuk No Ktp dan umur adalah integer. Tetapi properti dari nilai atribut dapat berbeda. Misalnya No Ktp tidak memiliki limit tetapi umur memiliki nilai maksimum dan minimum

Tipe Atribut berdasarkan sifatnya

Tipe atribut		Deskripsi	Contoh
Kategori (kualitatif)	Nominal	Nilai dari atribut nominal adalah nama-nama yang berbeda, yaitu nilai nominal hanya menyediakan informasi yang cukup untuk membedakan satu objek dengan objek yang lain. (= dan ≠)	Kode pos, No KTP, no induk mahasiswa, jenis kelamin
	Ordinal	Nilai dari atribut ordinal menyediakan informasi yang cukup mengurutkan objek. (<, >)	Predikat kelulusan
Numerik (Kuantitatif)	Interval	Untuk atribut interval, perbedaan antarnilai adalah sesuatu yang berarti, adanya unit pengukuran. (+,–)	Suhu dalam Celcius
	Ratio	Untuk variabel rasio, perbedaan dan rasio merupakan hal yang berarti. (*, /)	umur, panjang, tinggi

Atribut Diskret dan Kontinyu

Atribut Diskret

- Jika mempunyai nilai dalam himpunan jumlah yang terbatas
- Contoh : Kode pos, Jenis Kelamin, suhu
- Seringkali direpresentasikan sebagai variable integer
- Atribut biner adalah kasus spesial dari atribut diskret
- Biasanya ditemui pada atribut kategoris

Atribut Kontinyu

- Memiliki jangkauan nilai real
- Contoh : tinggi badan, berat badan
- Biasanya menggunakan *floating point*. Tetapi ukuran presisi jumlah angka di belakang koma tetap digunakan

Latihan

- Klasifikasikan atribut berikut sebagai atribut biner, diskret, atau kontinyu. Kemudian klasifikasikan atribut tersebut sebagai atribut kualitatif (nominal atau ordinal) atau kuantitatif (interval atau rasio). Contoh: umur dalam tahun, jawaban: diskret, kuantitatif, rasio.
- Dalam beberapa kasus, mungkin terdapat atribut yang dapat dikelompokkan ke lebih dari 2 tipe.
  - Waktu dalam AM atau PM
  - Sudut dalam derajat (antara 0 dan 360 derajat)
  - Jumlah pasien dalam sebuah rumah sakit
  - Nomor ISBN dari sebuah buku (Contoh format ISBN: **0-07-144373-8**)
  - Jarak ruang kuliah dari kantor pusat di sebuah Universitas
  - Medali emas, perak dan perunggu yang diberikan dalam sebuah kejuaraan

Tipe Set Data

- **Data Rekord**
  - Data Matrix
  - Data Dokumen
  - Data Transaksi
- **Data Grafik**
  - World Wide Web
  - Molecular Structures
- **Data Terurut**
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

Data Record

- Data yang terdiri dari kumpulan baris data (*records / entries / objects*), dimana setiap barisnya terdiri dari sejumlah atribut yang tetap.

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

Data Record (Data Transaksi)

- Data transaksi merupakan tipe spesial dari data rekord, dimana
  - Setiap rekord (transaksi) mengandung sekumpulan item

Contoh, data keranjang belanja dari sebuah supermarket.
Data transaksinya berisi kumpulan item dan jumlah item untuk setiap transaksi bisa berbeda dengan transaksi lainnya.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

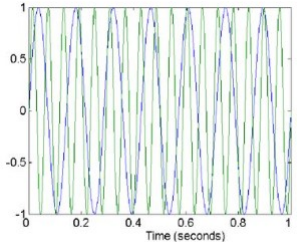
Kualitas Data

- Apa jenis masalah dari kualitas data?
- Bagaimana kita dapat mendeteksi masalah dalam data?
- Apa yang dapat kita lakukan untuk menghadapi masalah tersebut?

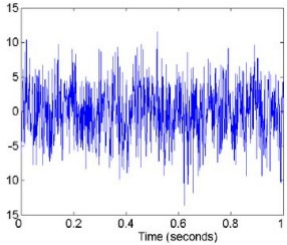
- Contoh dari masalah kualitas data:
  - Noise dan outlier
  - Missing Value
  - Duplicate data

Noise

- Noise mengarah kepada terjadinya modifikasi dari nilai yang sebenarnya.
- **Contoh** :
  - Penyimpangan dari suara seseorang ketika berbicara dengan menggunakan jaringan sinyal telepon yang jelek



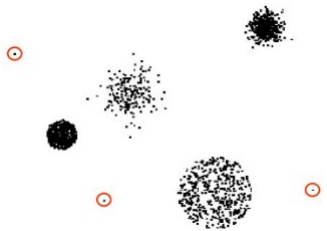
Two Sine Waves



Two Sine Waves + Noise

Outlier

- Outlier adalah objek data dengan karakteristik berbeda dari karakteristik sebagian besar objek pada set data.



Missing Value

- **Kenapa bisa ada missing value?**
  - Datanya tidak dapat diperoleh
    - (contoh : orang mungkin menolak untuk memberitahu umur dan berat badannya)
  - Atribut mungkin tidak dapat diaplikasikan untuk semua kasus
    - (contoh : pendapatan tahunan tidak dapat diaplikasikan ke anak-anak)
- **Menangani Missing Value**
  - Eliminasi objek data tersebut
  - Estimasi nilai dari missing value
  - Abaikan missing value tersebut selama proses analisis
    - Misalkan objek tersebut akan digunakan pada proses clustering. Jarak kedekatan yang diperlukan dalam proses clustering dapat dihitung dengan menggunakan atribut lain yang tidak hilang

Duplicate Data

- Di dalam set data mungkin terdapat duplikasi objek data.
  - Biasanya terjadi ketika terjadi penggabungan data dari sumber yang berbeda
  - **Contoh** : Orang yang sama dengan banyak alamat email
- **Penghapusan Data**
  - Proses yang dilakukan untuk menangani masalah duplikasi data

Praproses Data

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation (Penggabungan)

- Menggabungkan dua atau lebih atribut (atau objek) menjadi satu atribut (atau objek)
- **Tujuannya adalah :**
  - Pengurangan Data
    - Mengurangi jumlah atribut atau objek
  - Perubahan skala
    - Kota digabungkan menjadi provinsi, negara, dll
  - Agar data lebih seimbang
    - Data yang digabungkan cenderung memiliki perubahan yang sedikit

Contoh Aggregation

Cabang	IDT	Tanggal	Total
Gresik	2012102	30-01-2012	250.000
Gresik	2012103	30-01-2012	300.000
Surabaya	2012201	30-01-2012	500.000
Surabaya	2012202	30-01-2012	450.000
Surabaya	2012203	31-01-2012	350.000

Cabang	Tanggal	Total
Gresik	30-01-2012	550.000
Surabaya	30-01-2012	950.000
Surabaya	31-01-2012	350.000

Sampling

- Sampling merupakan pendekatan yang umum digunakan untuk pemilihan bagian (subset) dari objek/data secara keseluruhan yang akan dianalisis.
- Alasan penggunaan sampling:
  - Penggunaan seluruh data membuat proses yang harus dilakukan algoritma data mining menjadi lama.

Sampling

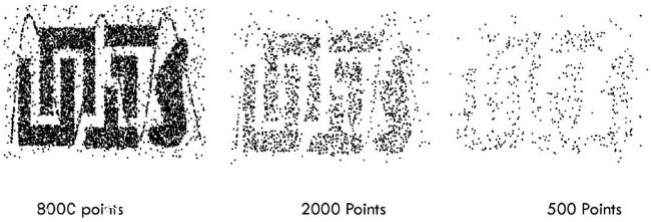
- Kunci utama dalam sampling:
  - Sampel data akan bekerja hampir sama dengan seluruh data jika sampel tersebut mampu mewakili (representatif) seluruh data.
  - Sampel disebut representatif jika diperkirakan mempunyai sifat yang sama dengan seluruh data.
  - Jika menggunakan rata-rata (mean) pada proses sampling, maka sebuah sampel dikatakan representatif jika sampel tersebut memiliki standard deviation yang mendekati data asli.

Tipe Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item.
  - Ada 2 jenis: Sampling tanpa pengembalian dan sampling dengan pengembalian.
- Sampling tanpa pengembalian
  - Setiap data yang sudah terambil untuk digunakan sebagai sampel tidak dikembalikan lagi ke data aslinya.
- Sampling dengan pengembalian
  - Setiap data yang terambil untuk sampel dikembalikan ke data asli.

Sample Size

- Ukuran sampel yang lebih besar meningkatkan peluang sampel tersebut menjadi sampel yang representatif, tetapi juga mengeliminasi banyak keuntungan dari proses sampling.
- Sebaliknya, dengan ukuran sampel yang lebih kecil, bentuk asli data mulai tidak tampak.



Dimensionality Reduction

- **Tujuan:**
  - Mengurangi penggunaan waktu dan memori yang dibutuhkan untuk eksekusi algoritma data mining.
  - Memungkinkan data untuk lebih mudah divisualisasikan.
  - Mungkin membantu untuk mengeliminasi fitur yang tidak relevan atau mengurangi noise.
- **Teknik:**
  - Principle Component Analysis

Feature Subset Selection

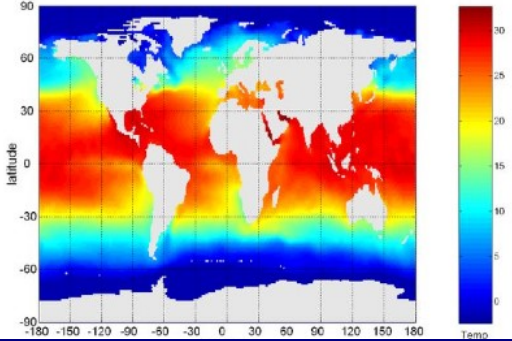
- Cara lain untuk mereduksi dimensi data.
- Fitur yang tidak relevan:
  - Tidak memiliki informasi yang berguna bagi tugas data mining yang sedang dikerjakan.
  - Contoh: Nomor induk mahasiswa tidak relevan dengan tugas memprediksi IPK mahasiswa.

Visualisasi

- Visualisasi adalah konversi dari data menjadi sebuah format visual atau tabular sehingga karakteristik data dan hubungan antar data atau atribut dapat dianalisis.
- Visualisasi dari data adalah salah satu teknik yang tepat untuk eksplorasi data.
  - Dapat mendeteksi pola umum dan trend data.
  - Dapat mendeteksi outlier dan pola yang tidak biasa.

Example: Sea Surface Temperature

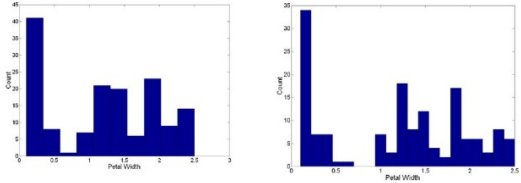
- The following shows the Sea Surface Temperature (SST) for July 1982.
- Tens of thousands of data points are summarized in a single figure.



Teknik Visualisasi: Histograms

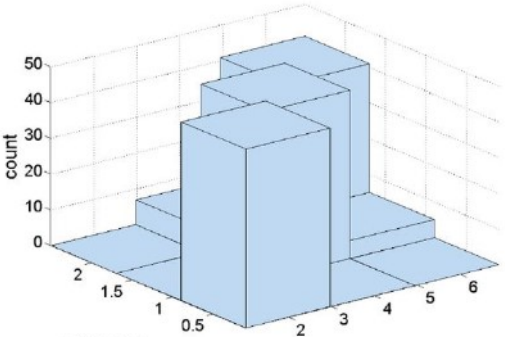
- **Histogram**
- Biasanya menunjukkan distribusi nilai dari sebuah single variable.
- Membagi nilai menjadi beberapa bagian.
- Tinggi dari setiap bar menunjukkan jumlah dari objek.

Example: Petal Width (10 and 20 bins, respectively)



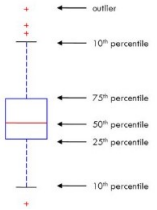
Two-Dimensional Histograms

- Menunjukkan distribusi gabungan dari dua atribut.
- Example: petal width and petal length



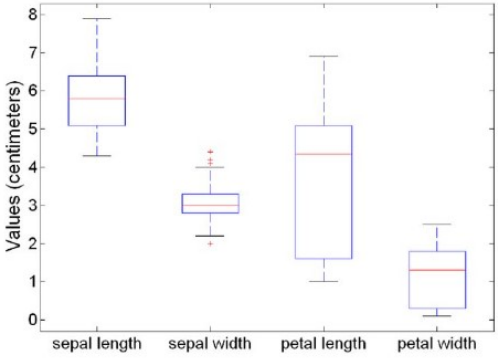
Teknik Visualisasi: Box Plots

- **Box Plots**
- Cara lain untuk menunjukkan distribusi dari data.



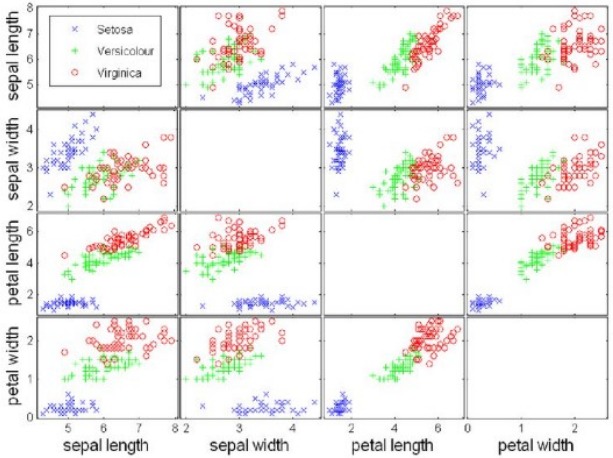
Example of Box Plots

- Box plots can be used to compare attributes.



Teknik Visualisasi: Scatter Plots

- **Scatter plots**
- Nilai atribut menjelaskan posisi.
- Scatter plot berguna untuk mendapatkan ringkasan data hubungan antara beberapa pasangan atribut.


Setosa
Versicolour
Virginica
sepal length
sepal width
petal length
petal width

Thank You!

Institut Informatika & Bisnis
DARMAJAYA
Yayasan Alfian Husin