Data Visualization

Data

Made Satria Wibawa, M.Eng.

DATA

- Data adalah kumpulan objek/poin beserta atributnya

- Atribut adalah karakteristik atau sifat dari sebuah objek
  - Contohnya : warna mata, suhu, dsb
  - Atribut juga disebut variabel, kolom, fitur
- Kumpulan dari atribut membentuk objek data
  - Objek juga disebut poin, record, baris, row, sample, instance

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

DATA

Tipe Atribut	Deskripsi	Contoh	Operasi
Nominal	Nilai dari atribut nominal hanya nama yang berbeda, yang hanya dapat menyediakan informasi untuk membedakan satu objek data dengan lainnya. (=, ≠)	Kode pos, ID karyawan, NIM, warna mata, gender {*male, female*}	mode, entropy, contingency correlation
Ordinal	Nilai dari atribut ordinal menyediakan informasi untuk mengurutkan objek data. (<, >)	{*good, better, best*}, IPK, nomor rumah	median, percentiles, rank correlation, run tests, sign tests
Interval	Pada atribut interval, selisih nilai antar objek data memiliki makna. (+, -)	tanggal, suhu dalam Celcius atau Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	Untuk atribut ratio, selisih dan rasio memieliki makna. Nilai 0 merupakan 0 absolut/tidak ada nilai yang dapat diukur (*, /)	suhu dalam Kelvin, nilai uang, jumlah, umur, bobot, panjang, arus listrik	geometric mean, harmonic mean, percent variation

NILAI ATRIBUT

1. Atribut Diskrit
   - Memiliki nilai terbatas (finite)
   - Contohnya kode pos, jumlah
   - Seringkali direpresentasikan dalam tipe integer
   - Atribut biner adalah atribut diskrit yang hanya memiliki dua nilai
2. Atribut Kontinyu
   - Memiliki nilai real
   - Contohnya suhu, bobot, panjang
   - Seringkali direpresentasikan dalam tipe float

DESKRIPSI STATISTIK

Deskripsi statistik dibagi menjadi dua, yaitu :

- Melihat titik pusat data
  - mean, median, mode
- Melihat sebaran data
  - range, quartiles, variance, standard deviation, interquartile range

TITIK PUSAT DATA

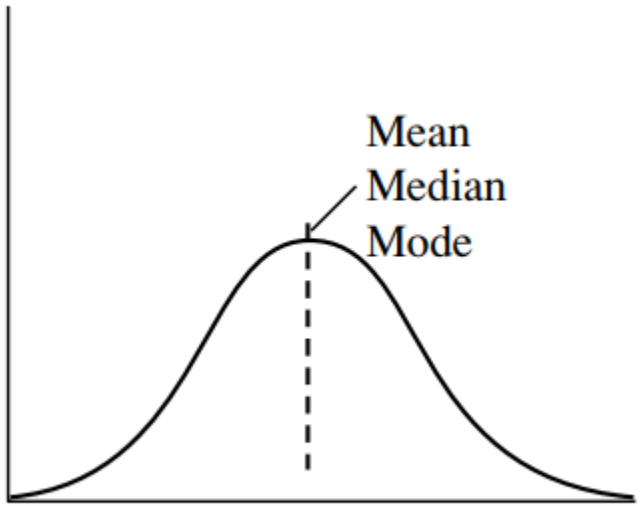
Misalkan kita memiliki atribut X, dengan objek x dari sekumpulan N buah data. Maka :

- mean : $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$

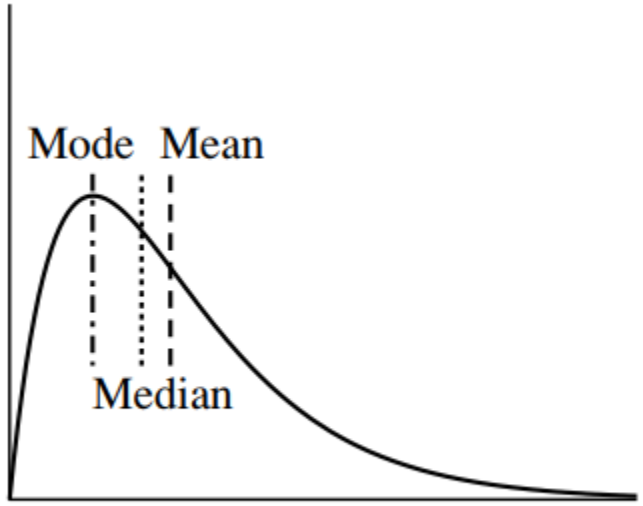
- median: odd $x_m = x_{\frac{n+1}{2}}$ even $x_m = \frac{x_{\frac{n}{2}} + x_{\frac{n+2}{2}}}{2}$

- mode : *most frequent data (unimodal, bimodal, trimodal, multimodal)*

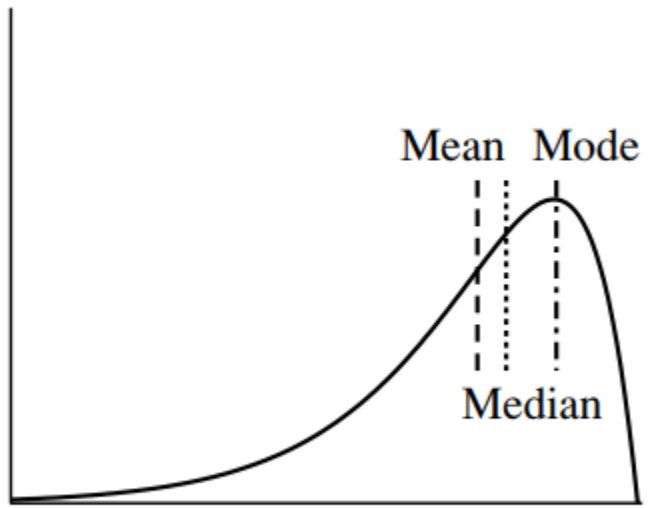
TITIK PUSAT DATA



(a) Symmetric data



(b) Positively skewed data



(c) Negatively skewed data

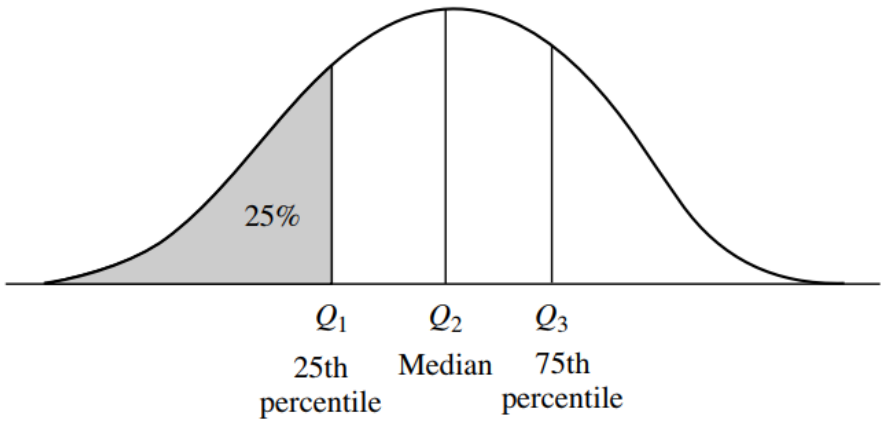
SEBARAN DATA

- **range** : $x_{max} - x_{min}$
- **quantile** : titik yang membagi data ke kumpulan bagian yang 'equal'

  jika dibagi menjadi 4 bagian, disebut quantile

  jika dibagi menjadi 100 bagian disebut percentile

  interquartile range ($\boldsymbol{IQR}$) merupakan jarak quartile 1 (Q_1) dan quartile 2 (Q_2)



SEBARAN DATA

- **standard deviation**

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

- **variance**

$$s = s^2$$

kedua parameter ini menyatakan sebaran data dalam satu atribut, jika nilainya besar maka nilai data semakin bervariasi/menyebar

KUALITAS DATA

- Permasalahan dalam data.
- Cara deteksi permasalahan.
- Penanggulangan permasalahan.

Data sangat rentan terhadap data yang noise, missing value, dan inconsistent karena ukurannya yang biasanya besar dan kemungkinan berasal dari berbagai sumber yang beragam.

KUALITAS DATA

Accuracy

Completeness

Consistency

PERMASALAHAN DI DATA

- Missing value
- Outlier
- High dimensionality

MISSING VALUES

Missing value adalah nilai yang hilang pada salah satu atau beberapa atribut dalam sebuah objek data

- Alasan missing value
  - Informasi tidak dikumpulkan

    (misalnya orang enggan memberikan informasi umur dan gajinya)
  - Atribut yang tidak dapat diaplikasikan ke semua kasus

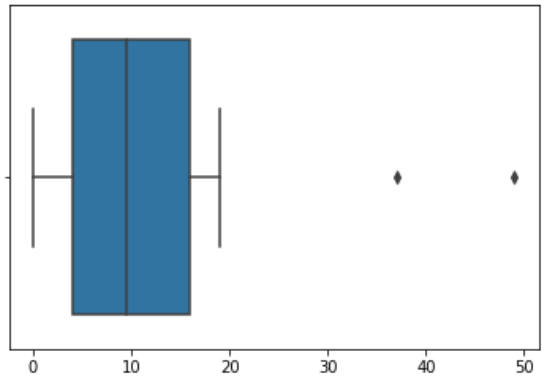
    (gaji tidak dapat diterapkan kepada anak-anak)
- Penanganan missing value
  - Hapus objek data
  - Perkirakan nilai missing value
  - Abaikan missing value

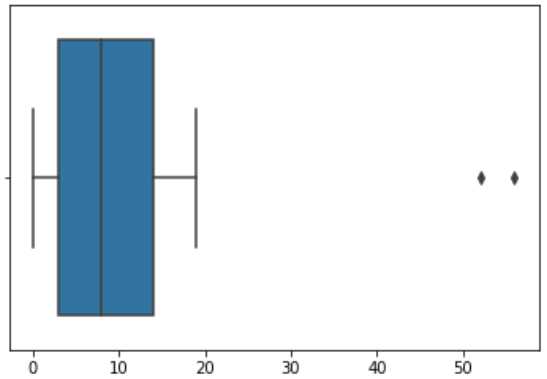
Index	Age	Sex	Income
1	NA	M	NA
2	39	NA	75000
3	NA	NA	NA
4	28	F	50000
...	...	...	...
10000	18	F	NA

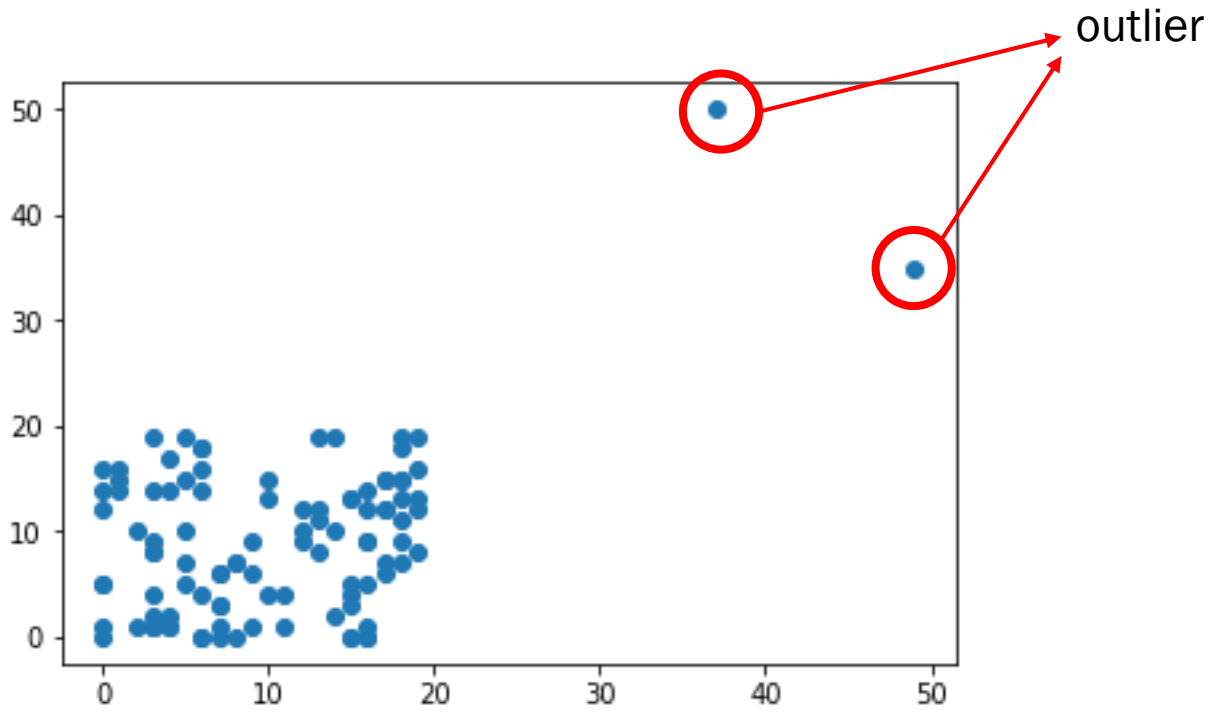
OUTLIER

Objek data yang memiliki karakteristik sangat jauh berbeda dari data kebanyakan

0	1
10	5
19	7
9	2
11	19
10	18
...	...
6	7
5	11
16	13
19	13
42	45



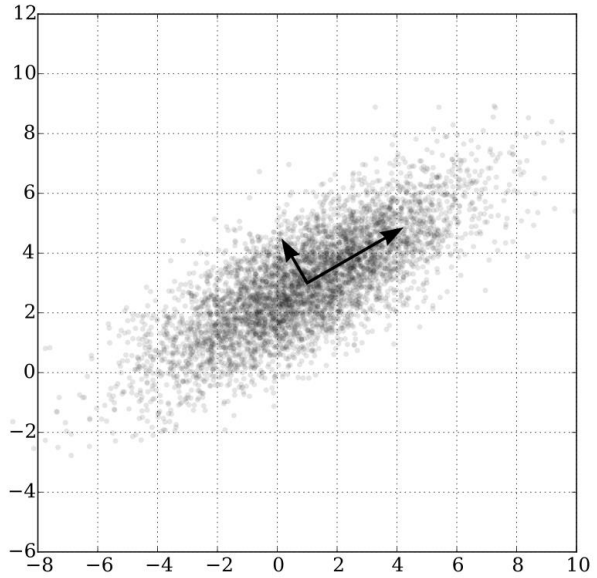




DATA BERDIMENSI TINGGI

Data berdimensi tinggi adalah data dengan jumlah variabel/atribut yang banyak, sehingga menyulitkan untuk visualisasi

- Penanganan
  - Principle Component Analysis
  - Singular Value Decomposition



DATA AGGREGATION

Merupakan proses untuk mengumpulkan dan merepresentasikan data dalam bentuk yang ringkas.

DATA AGGREGATION

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Total sales	19,795	23,005	31,711	40,728	50,440	60,953	74,143	93,321	120,312
Male	12,534	16,452	19,362	24,726	28,567	31,110	39,001	48,710	61,291
Female	7,261	6,553	12,349	16,002	21,873	29,843	35,142	44,611	59,021
Regular	9,929	14,021	17,364	20,035	27,854	34,201	36,472	52,012	60,362
Decaf	6,744	6,833	10,201	13,462	17,033	19,921	21,094	23,716	38,657
Mocha	3,122	2,151	4,146	7,231	5,553	6,831	16,577	17,593	21,293

Factoid
adalah sebuah fakta

Contoh : 36.7% kopi di tahun 2000 dikonsumsi oleh wanita

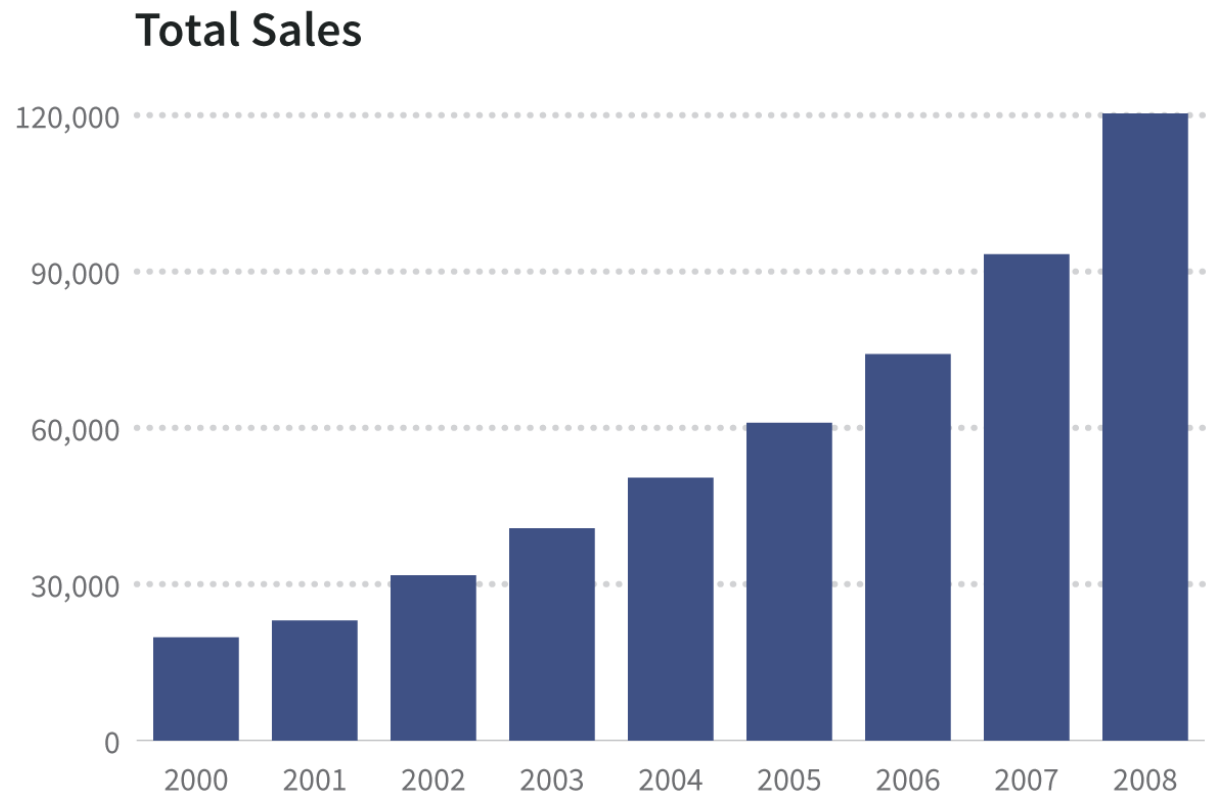
DATA AGGREGATION

Series

merupakan tipe informasi dimana variabel satu dibandingkan dengan variabel lainnya (biasanya variabel independent dengan variabel dependent). Seringkali variabel independent adalah waktu

Contoh :

Year	2000	2001	2002	2003
Total sales	19,795	23,005	31,711	40,728



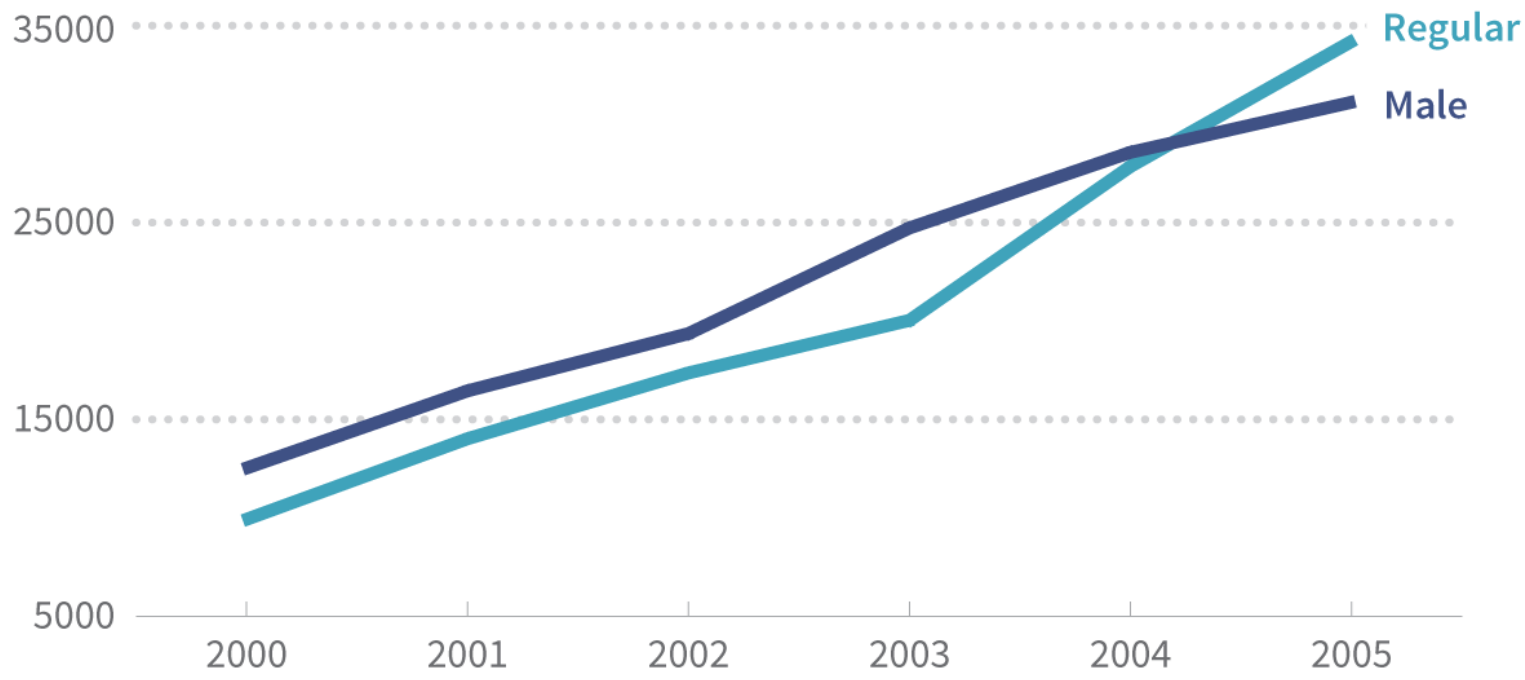
DATA AGGREGATION

Multiseries

memiliki beberapa informasi (variabel) dependent dan satu informasi (variabel) independent

Contoh :

Year	2000	2001	2002	2003	2004	2005
Male	12,534	16,452	19,362	24,726	28,567	31,110
Regular	9,929	14,021	17,364	20,035	27,854	34,201



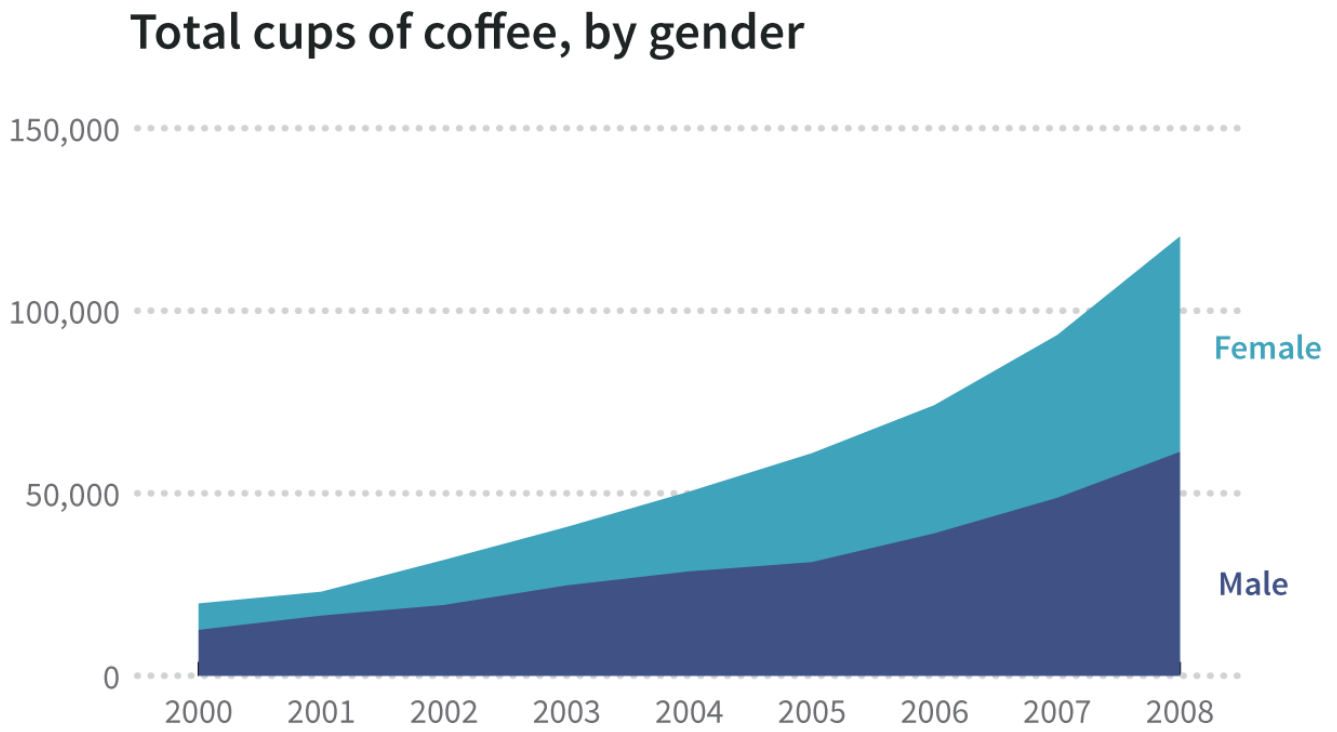
DATA AGGREGATION

Summable Multiseries

merupakan representasi data yang dibagi menjadi beberapa kelompok

Contoh :

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Male	12,534	16,452	19,362	24,726	28,567	31,110	39,001	48,710	61,291
Female	7,261	6,553	12,349	16,002	21,873	29,843	35,142	44,611	59,021



DATA AGGREGATION

Summary Records

merupakan representasi data yang dibagi menjadi beberapa kelompok

Contoh :

Name	Gender	Regular	Decaf	Mocha	Total
Bob Smith	M	2	3	1	6
Jane Doe	F	4	0	0	4
Dale Cooper	M	1	2	4	7
Mary Brewer	F	3	1	0	4
Betty Kona	F	1	0	0	1
John Java	M	2	1	3	6
Bill Bean	M	3	1	0	4
Jake Beatnik	M	0	0	1	1
Totals	5M, 3F	16	8	9	33

Apakah wanita memiliki jenis kopi tertentu yang digemari

Labels	Average of Regular	Average of Decaf	Average of Mocha
F	2.67	0.33	0.00
M	2.00	1.75	2.00
Grand Total	2.29	1.14	1.14

DATA AGGREGATION

Individual Transactions

merupakan rekaman dari transaksi

Contoh :

Timestamp	Name	Gender	Coffee
17:00	Bob Smith	M	Regular
17:01	Jane Doe	F	Regular
17:02	Dale Cooper	M	Mocha
17:03	Mary Brewer	F	Decaf
17:04	Betty Kona	F	Regular
17:05	John Java	M	Regular
17:06	Bill Bean	M	Regular
17:07	Jake Beatnik	M	Mocha
17:08	Bob Smith	M	Regular
17:09	Jane Doe	F	Regular
17:10	Dale Cooper	M	Mocha
17:11	Mary Brewer	F	Regular
17:12	John Java	M	Decaf
17:13	Bill Bean	M	Regular

Transaksi ini dapat dikumpulkan (agregasi) oleh kolom manapun. Timestamp dapat dikumpulkan ke dalam satuan (setiap jam, harian, atau tahunan). Dataset di awal dihasilkan dari data mentah ini, dengan dirangkum secara signifikan.

Q & A