University of Colorado
Boulder

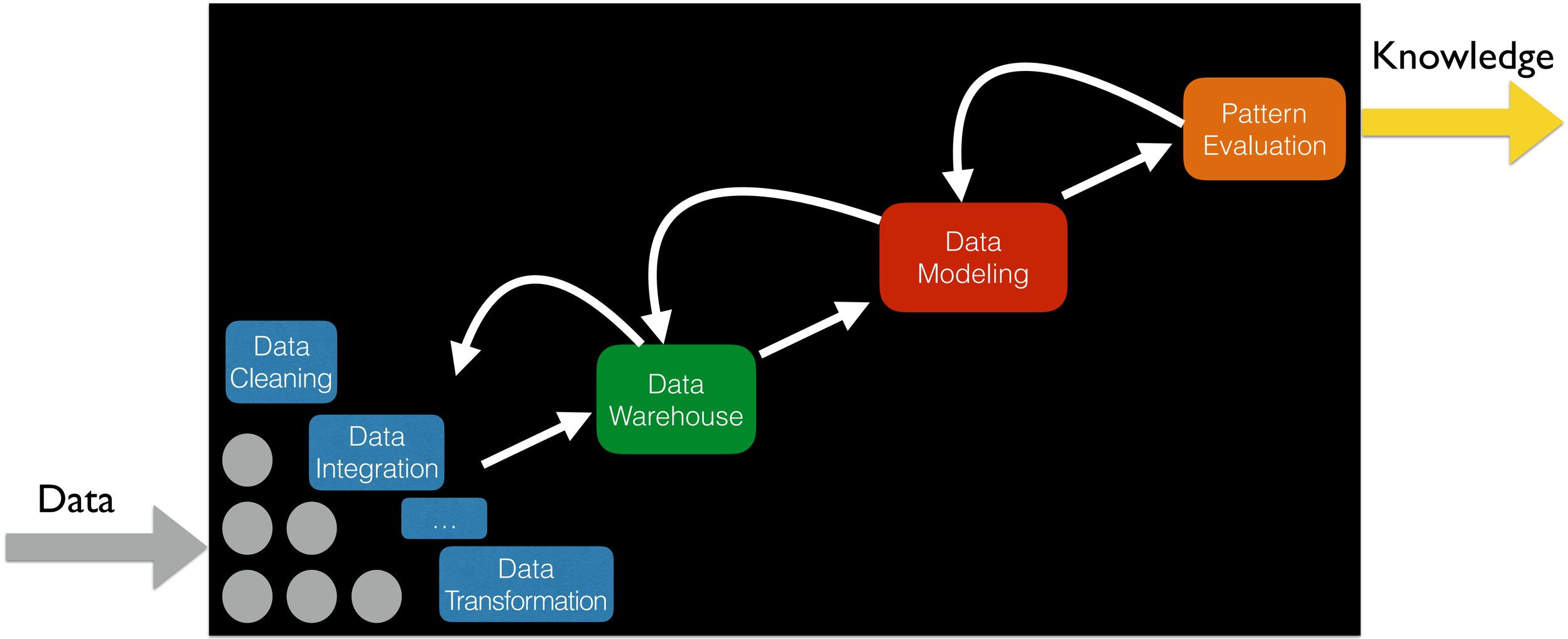
CSCI 4502/5502
Data Mining

Fall 2020
Lecture 03 (Sep 1)

Review

- Chapter 1: Introduction to data mining
  - data mining: discover interesting patterns in huge amounts of data
  - data mining pipeline
  - different views: data, knowledge, method, application
  - measure of pattern interestingness
  - major issues in data mining

Data Mining Pipeline



University of Colorado
Boulder

Chapter 2: Getting to Know Your Data

Chapter 2: Getting to know your data

- data objects and attribute types
- basic statistical description of data
- data visualization
- measuring data similarity and dissimilarity

Data Objects & Attributes

- Data set: a set of data objects
  - e.g., students, courses, customers, products
- Data object
  - an entity with certain attributes/features/dimensions/variables
  - e.g., patient_id, name, DOB, address, office visits, lab tests
- Attribute type: nominal, binary, ordinal, numeric

Attribute Types

- Nominal (categorical): e.g., major, occupation, city
- Binary (boolean, symmetric or asymmetric)
  - e.g., CS major? professor? Boulder?
- Ordinal: degree, professional rank, vehicle size class
- Numeric (quantitative)

Numeric Attributes

- Interval-scaled
  - e.g., 50 or 100 Fahrenheit degree; Year 2000 or 2020
- Ratio-scaled (true zero-point)
  - e.g., age, dollars, number of books, number of cars
- Discrete vs. continuous
  - discrete: finite or countably infinite; integers vs. real numbers

Chapter 2

- Getting to know your data
  - data objects and attribute types
  - basic statistical description of data
  - data visualization
  - measuring data similarity and dissimilarity

Statistical Description of Data

- Motivation: better understanding of the data
  - e.g., sales, traffic volume, #likes
- Basics: N, min, max
- Central tendency: mean, median, mode, midrange
- Dispersion: quartiles, interquartile range, variance

Central Tendency (1)

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

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$$

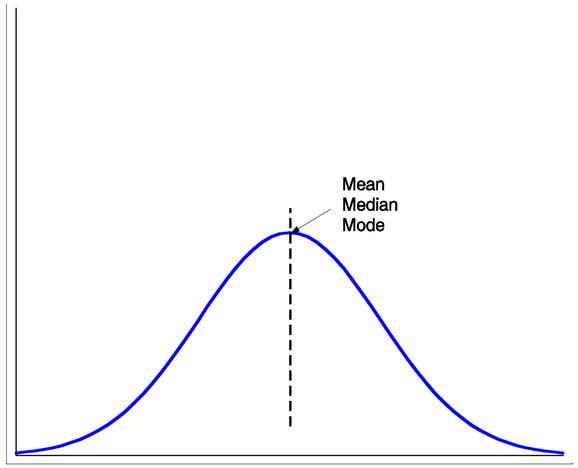
  - trimmed mean: chopping extreme values

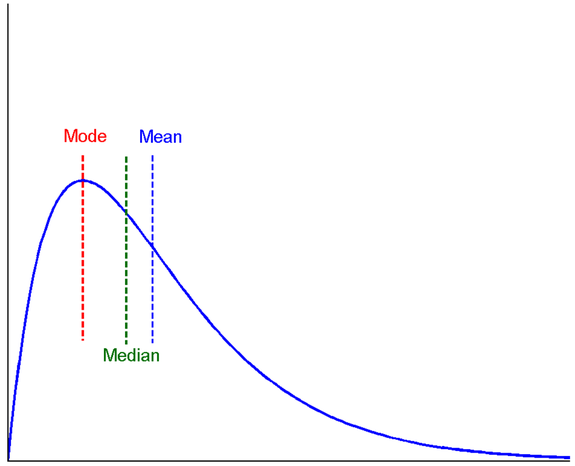
- **Median**
  - middle value if N is odd, otherwise
  - average of the middle two values

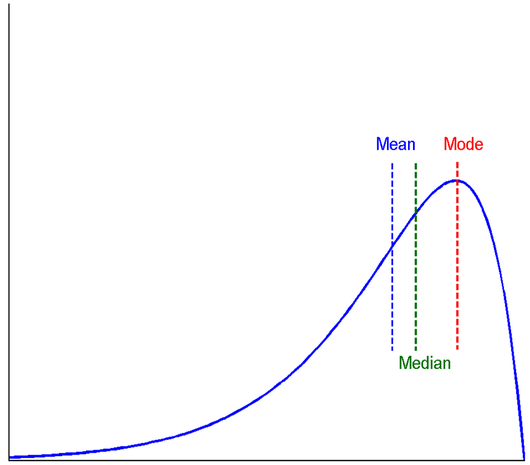
$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}}\right) width$$

Central Tendency (2)

- Mode: value that occurs most frequently
  - unimodal, biomodal, trimodal, multimodal
- Midrange: avg. of min and max

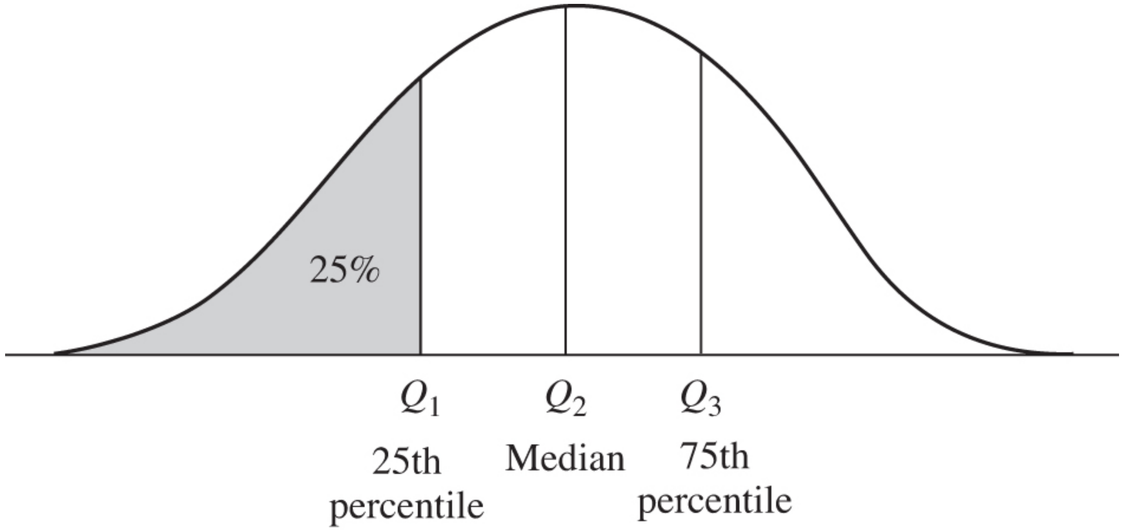






Data Dispersion (1)

- How much numeric data tend to spread
- Range: difference between max and min
- Quartiles: Q1 (25th percentile), Q3 (75th)
- Interquartile range
  - IQR = Q3 - Q1



Data Dispersion (2)

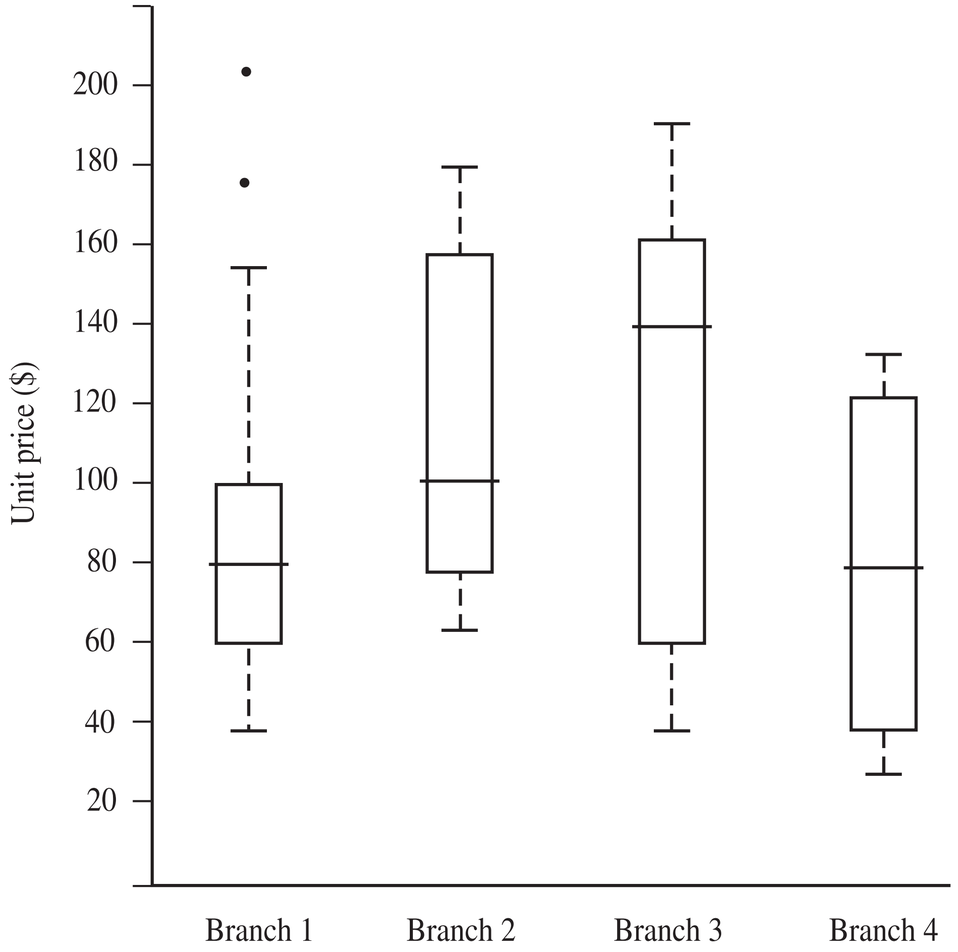
- Five number summary: min, Q1, median, Q3, max
- Outlier: value higher/lower than 1.5 x IQR of Q3/Q1
- Variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right]$$

- Standard deviation: square root of variance

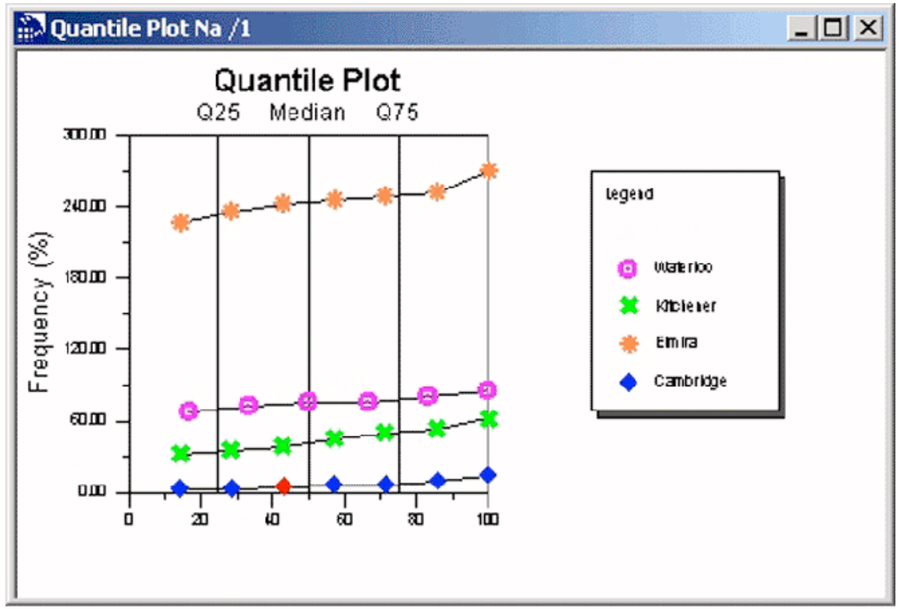
Data Dispersion (3)

- Boxplot
  - box: Q1, M, Q3, IQR
  - whiskers:
    - min, max, 1.5 x IQR
  - outliers



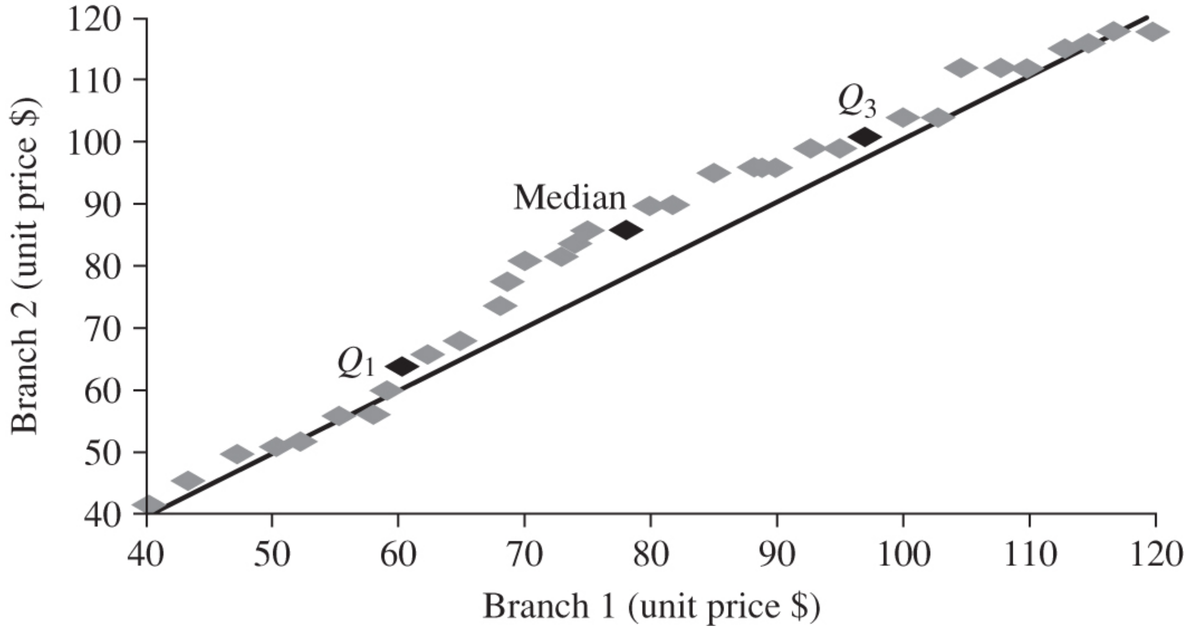
Graphic Displays (1)

- Quantile plot



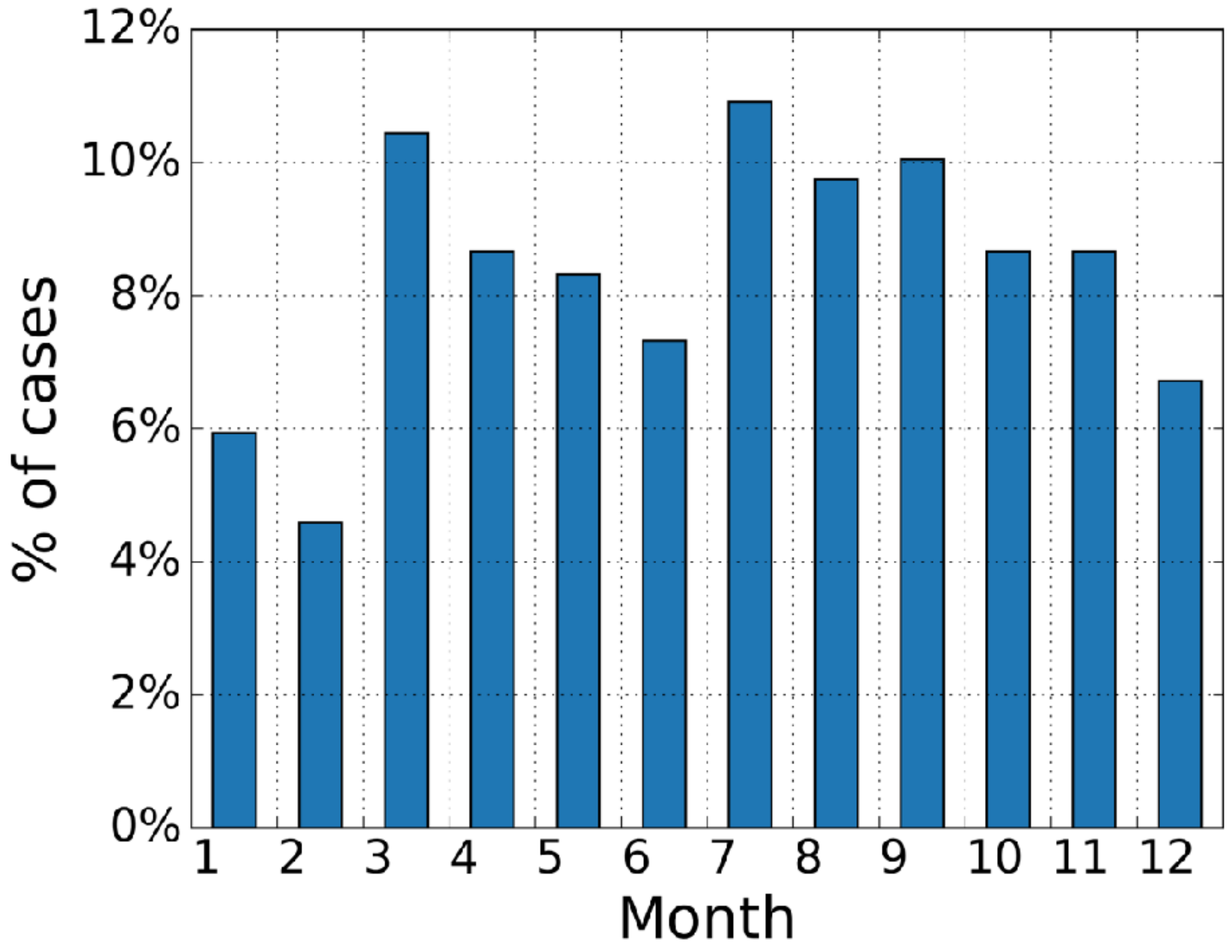
http://www.rockware.com/assets/products/70/features/114/230/aquachemplot10b.gif

- Quantile-quantile plot (Q-Q plot)



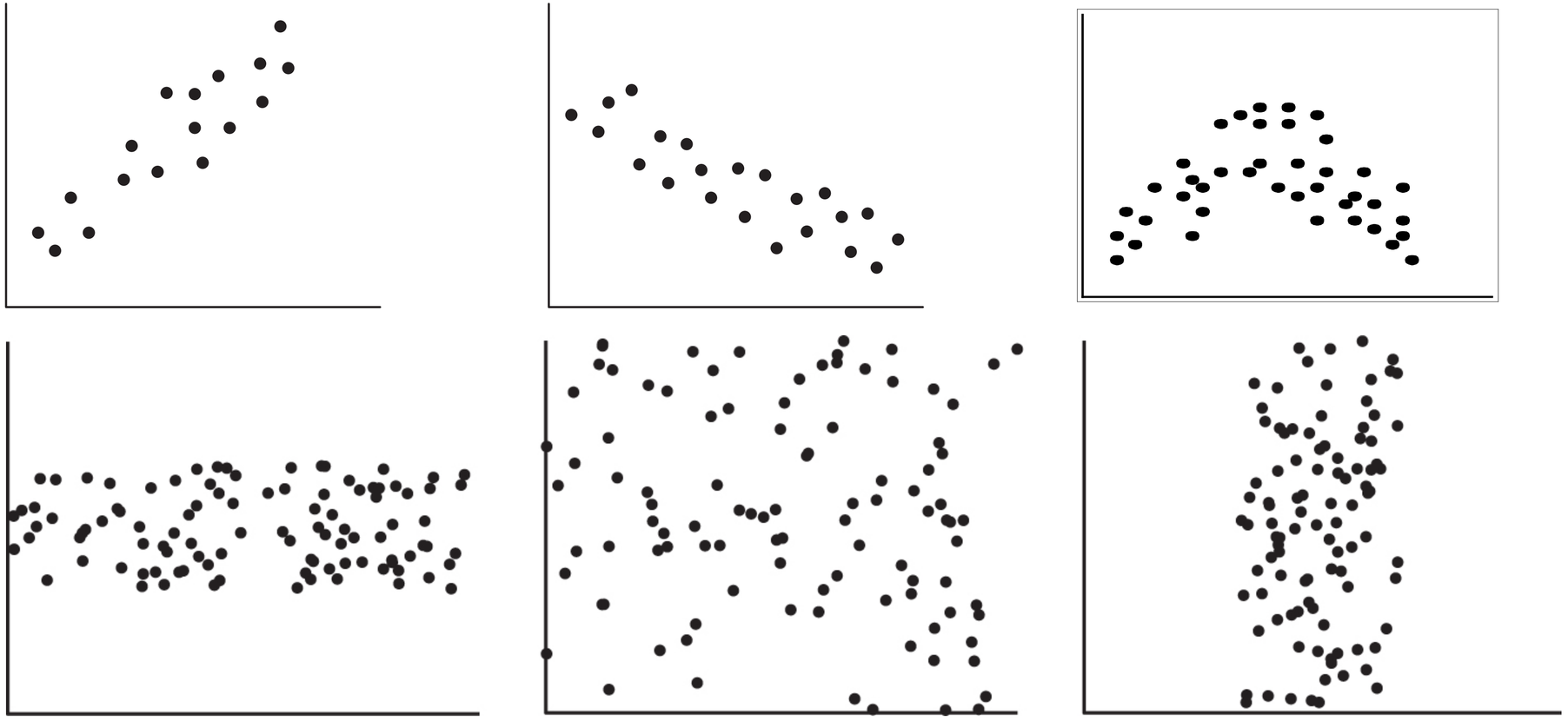
Graphic Displays (2)

- Histogram



Graphic Display (3)

- Scatter plot

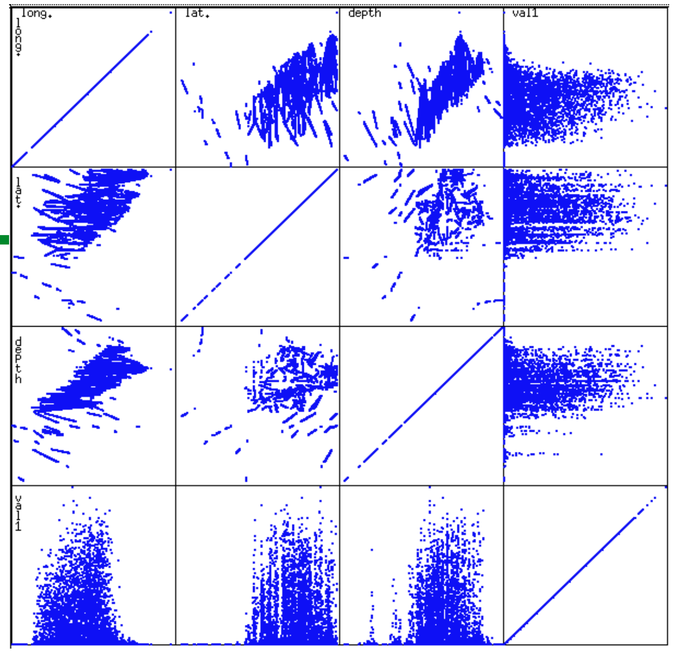


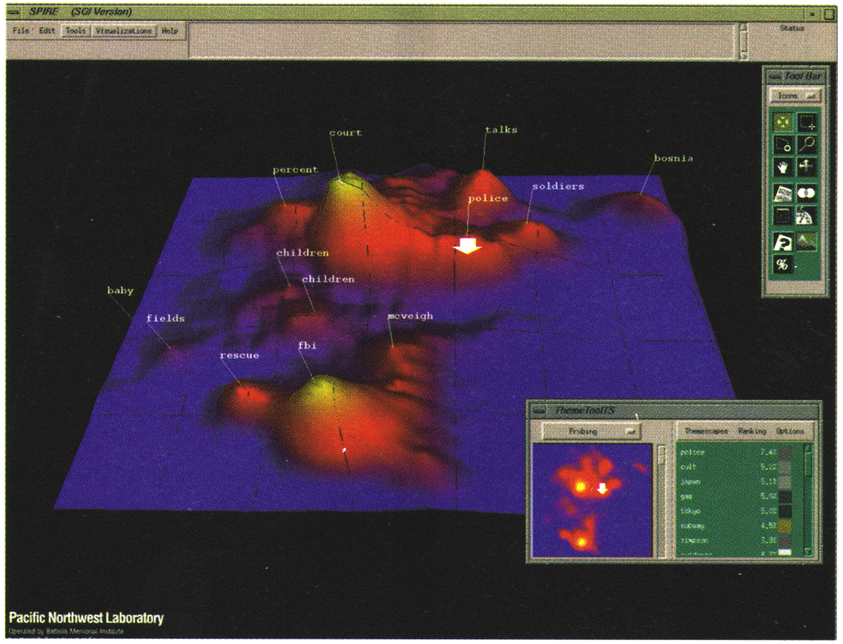
Chapter 2

- Getting to know your data
  - data objects and attribute types
  - basic statistical description of data
  - data visualization
  - measuring data similarity and dissimilarity

Data Visualization

- Why data visualization?
  - gain insights, qualitative overview, explore
- Visualization methods
  - pixel-oriented, icon-based, hierarchical
  - geometric projection
  - visualizing complex data and relations





https://www.1point21gws.com/insights/wp-content/uploads/2019/07/data.png

Chapter 2

- Getting to know your data
  - data objects and attribute types
  - basic statistical description of data
  - data visualization
  - measuring data similarity and dissimilarity

Data matrix

- object-by-attribute
- two modes

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity matrix

- object-by-object
- one mode

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Object Similarity/Dissimilarity

- Usually measured by distance
- Minkowski distance (L_p norm)

$$d(i,j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p)^{1/p}$$

- Euclidean distance (L_2 norm)

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}$$

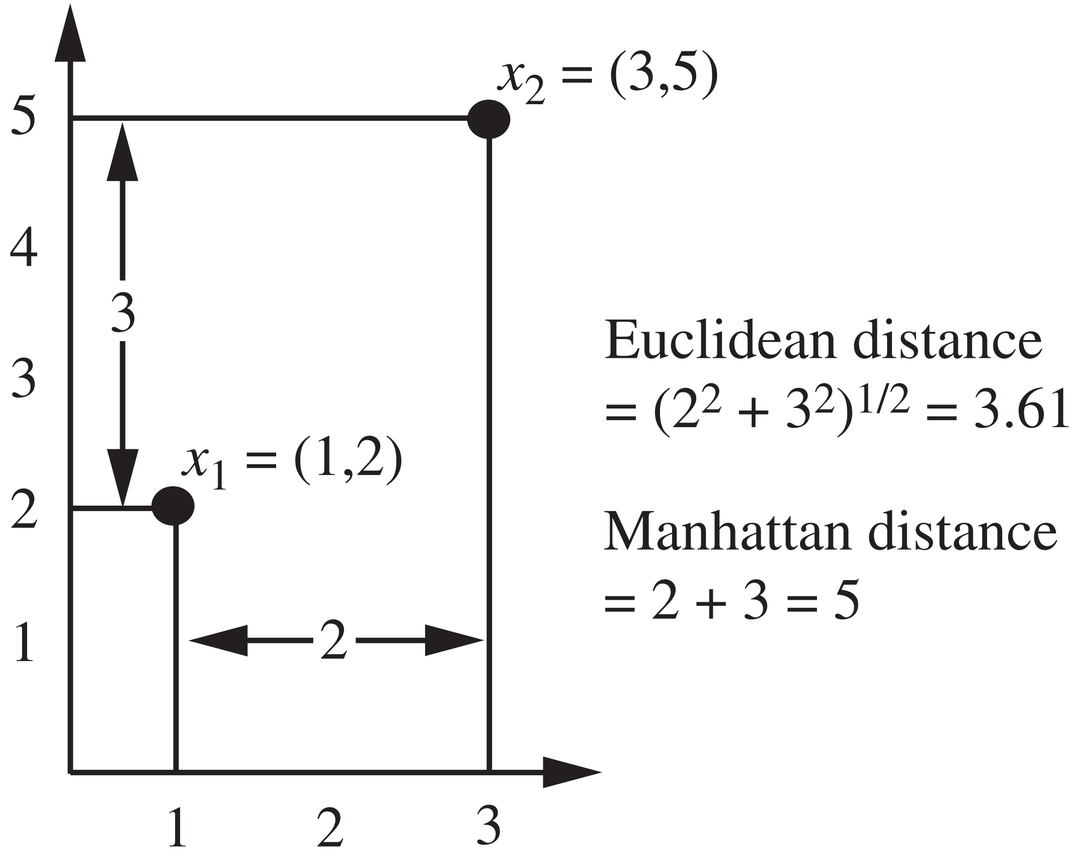
- Manhattan distance (L_1 norm)

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|$$

- Weighted distance

Distance Measure

- Euclidean distance vs. Manhattan distance
- Properties
  - d(i, j) >= 0
  - d(i, i) = 0
  - d(i, j) = d(j, i)
  - d(i, j) <= d(i, k) + d(k, j)
  - (triangular inequality)



Euclidean distance
$= (2^2 + 3^2)^{1/2} = 3.61$

Manhattan distance
$= 2 + 3 = 5$