University of Colorado Boulder

CSCI 4502/5502

Data Mining

Fall 2019
Lecture 03 (Sep 3)

Announcements

- Office hours
  - Tu 11am to 12pm; Fr 1pm to 2pm
  - ECCR 1B24, zoom, or by appointment
  - TA: Tu 4pm to 5pm, ECCR 1B10
- Homework 1
  - will be posted at moodle on Thursday
  - due at 9:30am, Thursday Sep 12
  - submit at moodle

Review

- Chapter 1: Introduction to Data Mining
  - Discovering interesting patterns in huge amounts of data
  - Knowledge discovery process
  - Different views of data mining
  - Measure of pattern interestingness
  - Major issues in data mining

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 Mining: Data View

- Database-oriented
- Sequence, temporal, time series data
- Spatial, spatial-temporal data
- Text, multimedia, Web data
- Graph and network data
- Data streams

- Structured, unstructured, semi-structured

Important Characteristics

- Dimensionality
  - curse of dimensionality
- Sparsity
  - only presence counts
- Resolution
  - patterns depend on the scale
- Distribution
  - centrality and dispersion

Data Objects & Attributes

- Data set: a set of data objects
  - e.g., students, courses, customers, products
- Data object
  - an entity with certain attribute(s)
    - also called feature, dimension, variable
    - e.g., student_id, name, DOB, major, address
- Attribute types
  - nominal, binary, ordinal, numeric

Attribute Types

- Nominal (categorical)
  - e.g., hair color, major, occupation
- Binary (boolean, symmetric or asymmetric)
  - e.g., gender, smoker, disease
- Ordinal
  - e.g., drink size, grade, professional rank
- Numeric (quantitative)

Numeric Attributes

- Interval-scaled
  - e.g., temperature 30 or 60 Celsius degree
  - e.g., Year 2000 or 2012
- Ratio-scaled (true zero-point)
  - e.g., words, dollars, age
- 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., wind speed, #Facebook friends
- Basics: N, min, max
- Central tendency
  - mean, median, mode, midrange
- Dispersion
  - quartiles, interquartile range, variance

Central Tendency (1)

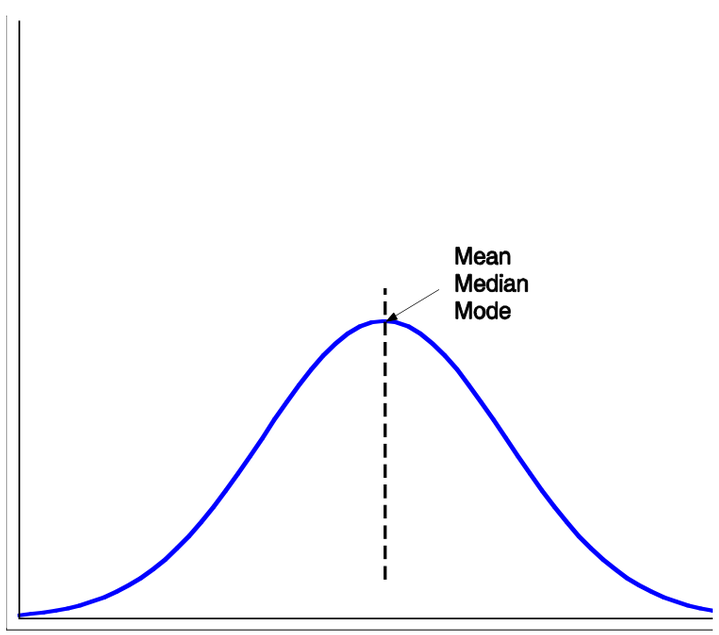
- Mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}$
  - weighted arithmetic mean

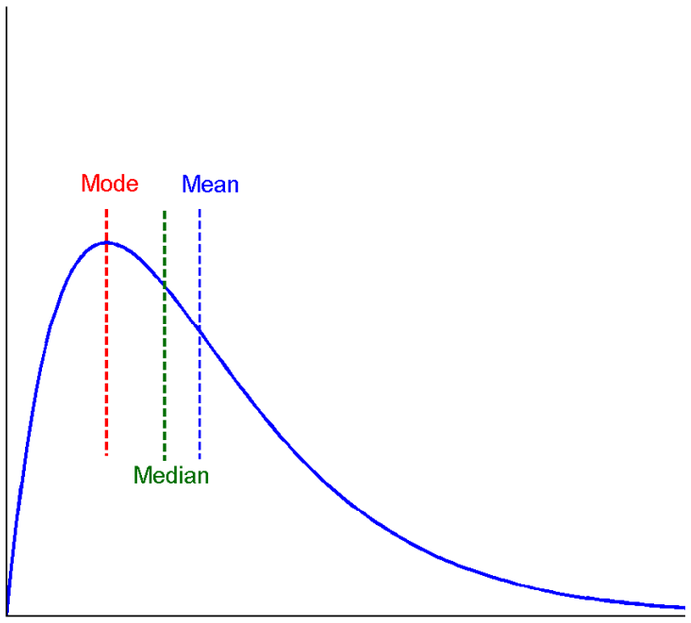
$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$$

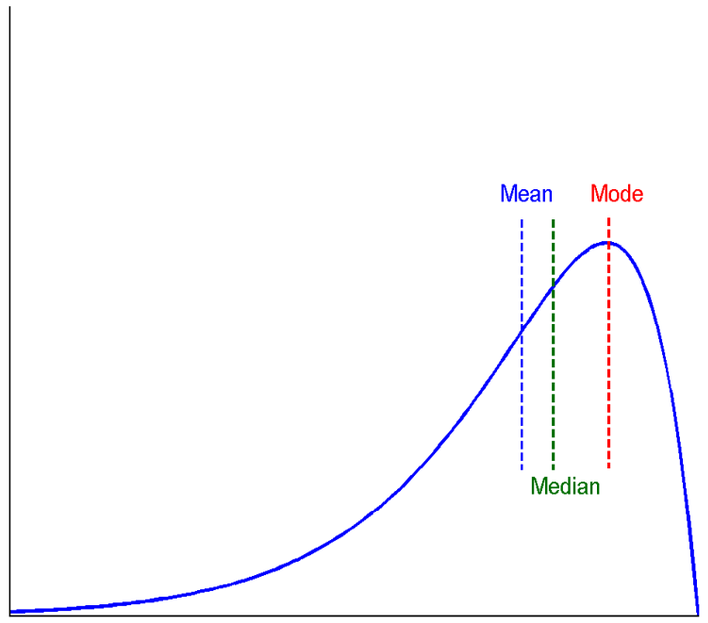
  - trimmed mean: chopping extreme values
- Median
  - middle value if N is odd, otherwise
  - average of the middle two values
  - estimation $median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$

Central Tendency (2)

- Mode: value that occurs most frequently
  - unimodal, biomodal, trimodal, multimodal
  - empirical $mean - mode = 3 \times (mean - median)$
- Midrange: avg. of min and max

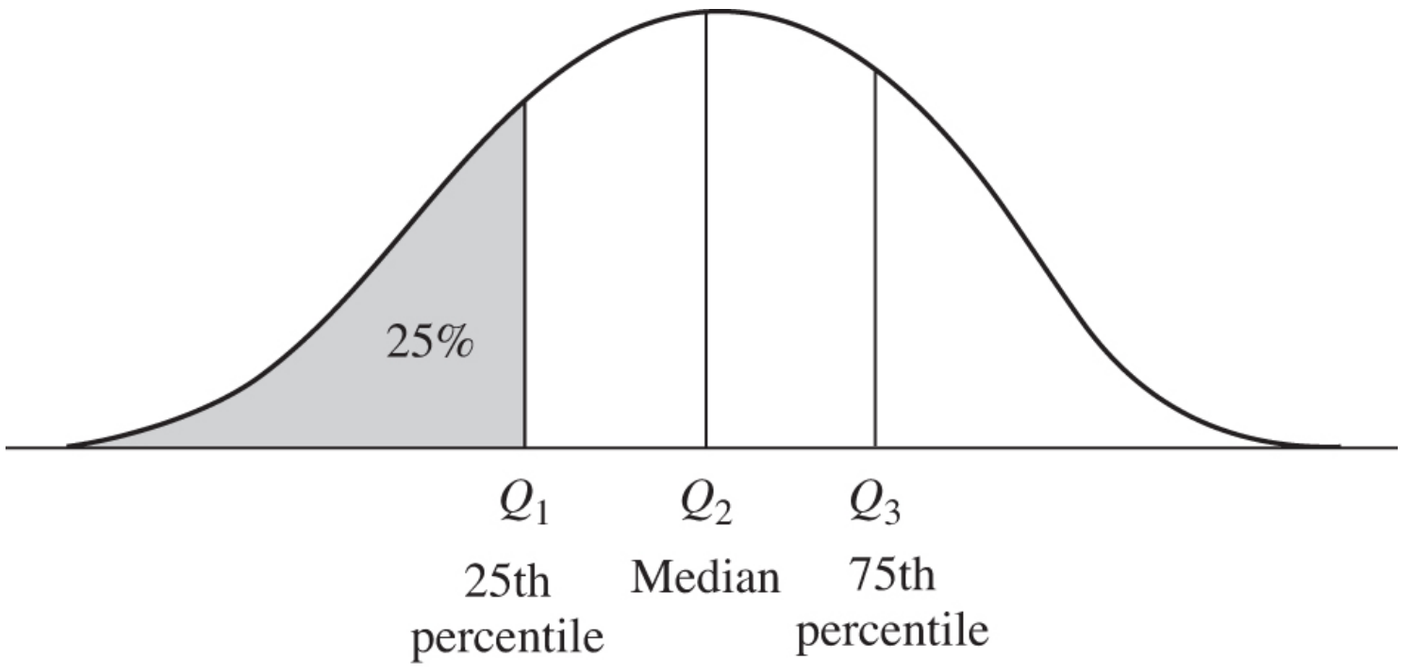






Data Dispersion (1)

- 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)

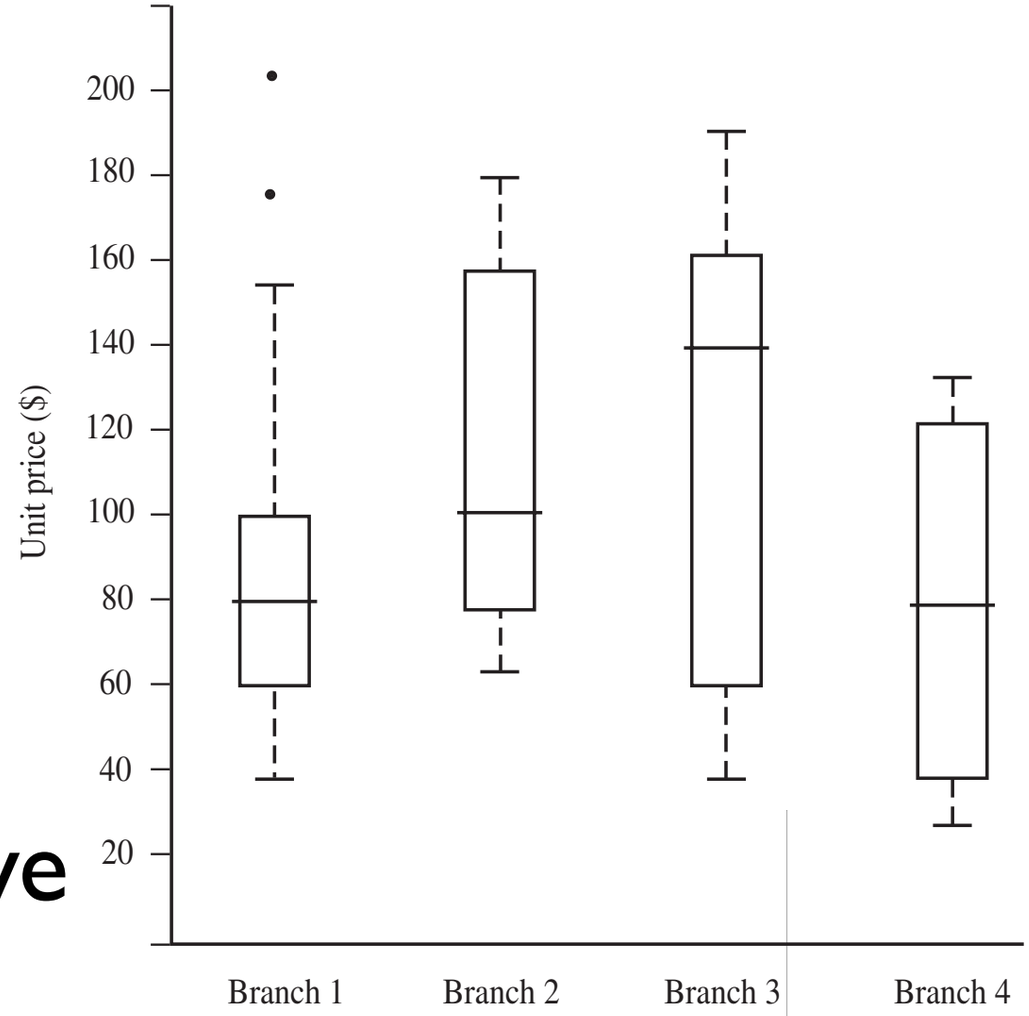
- Five number summary
  - min, Q1, median, Q3, max
- Outlier
  - value higher/lower than 1.5 x IQR of Q3/Q1
- Variance

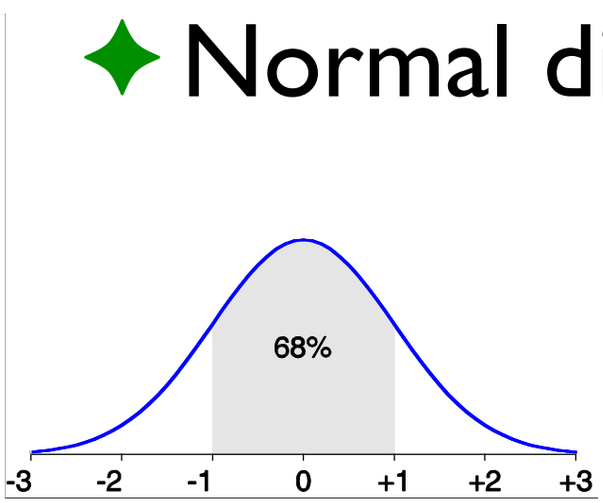
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right]$$

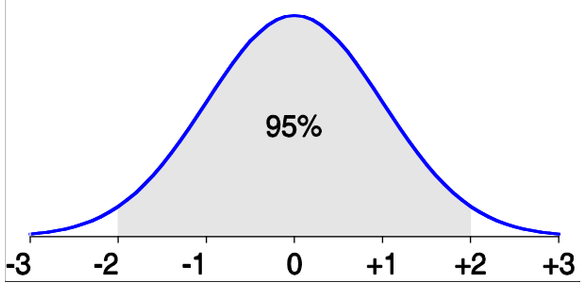
- Standard deviation
  - square root of variance

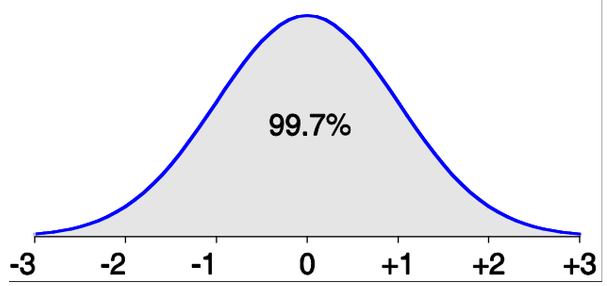
Data Dispersion (3)

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





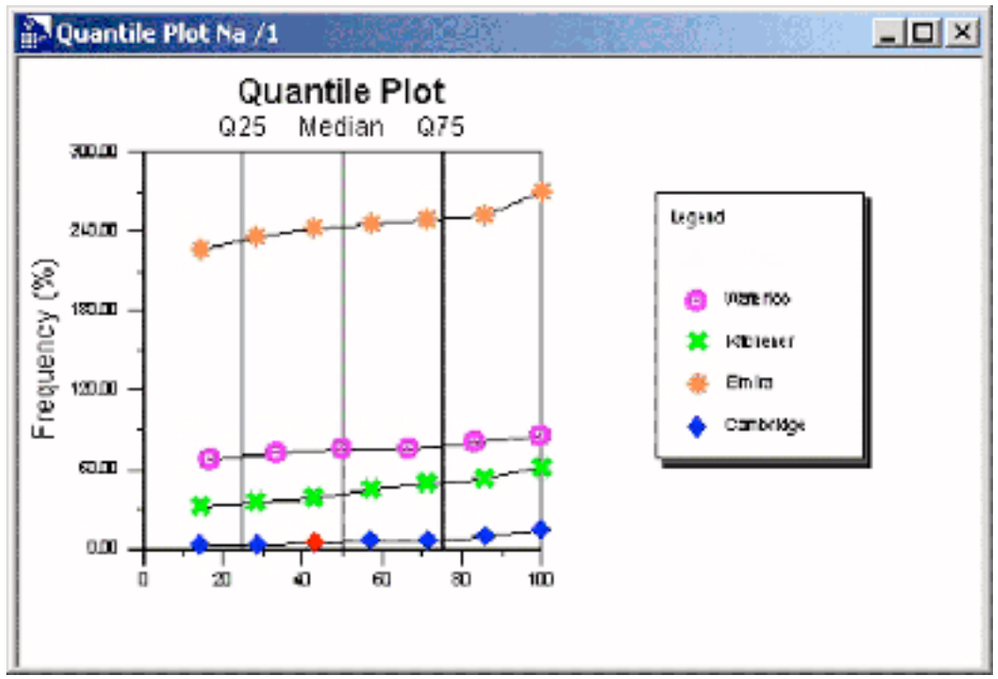




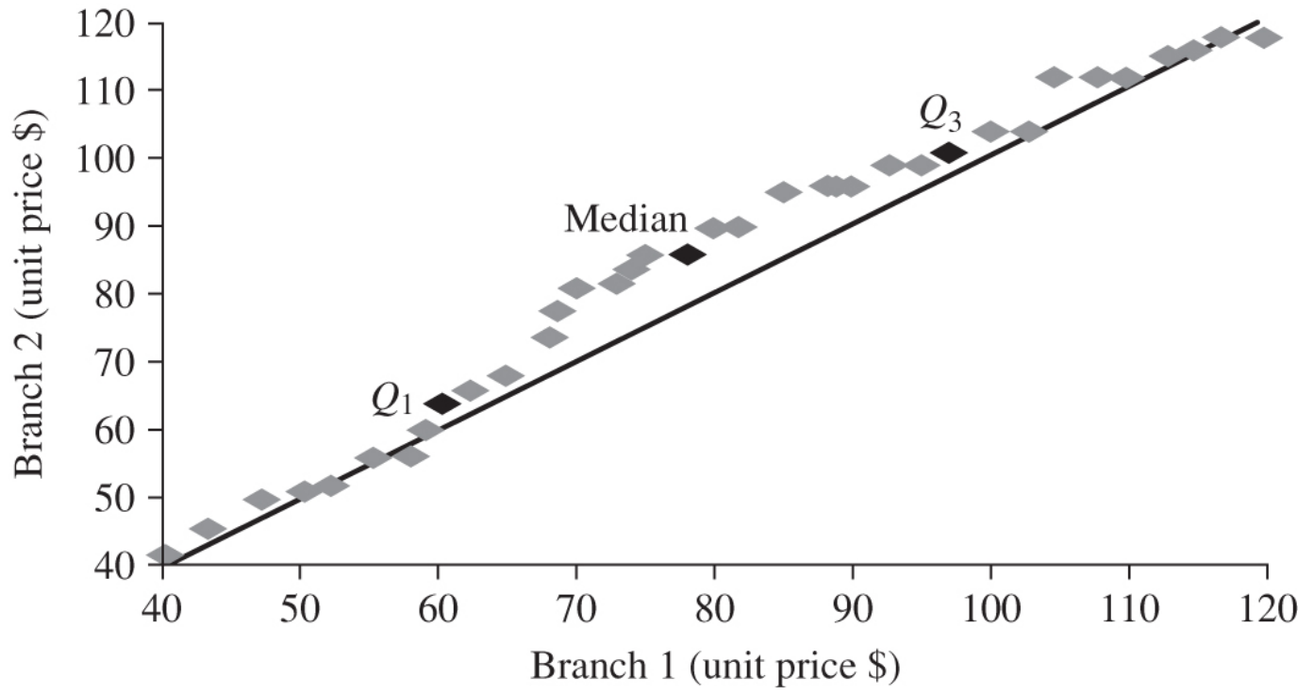
Graphic Displays (1)

- Quantile plot

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



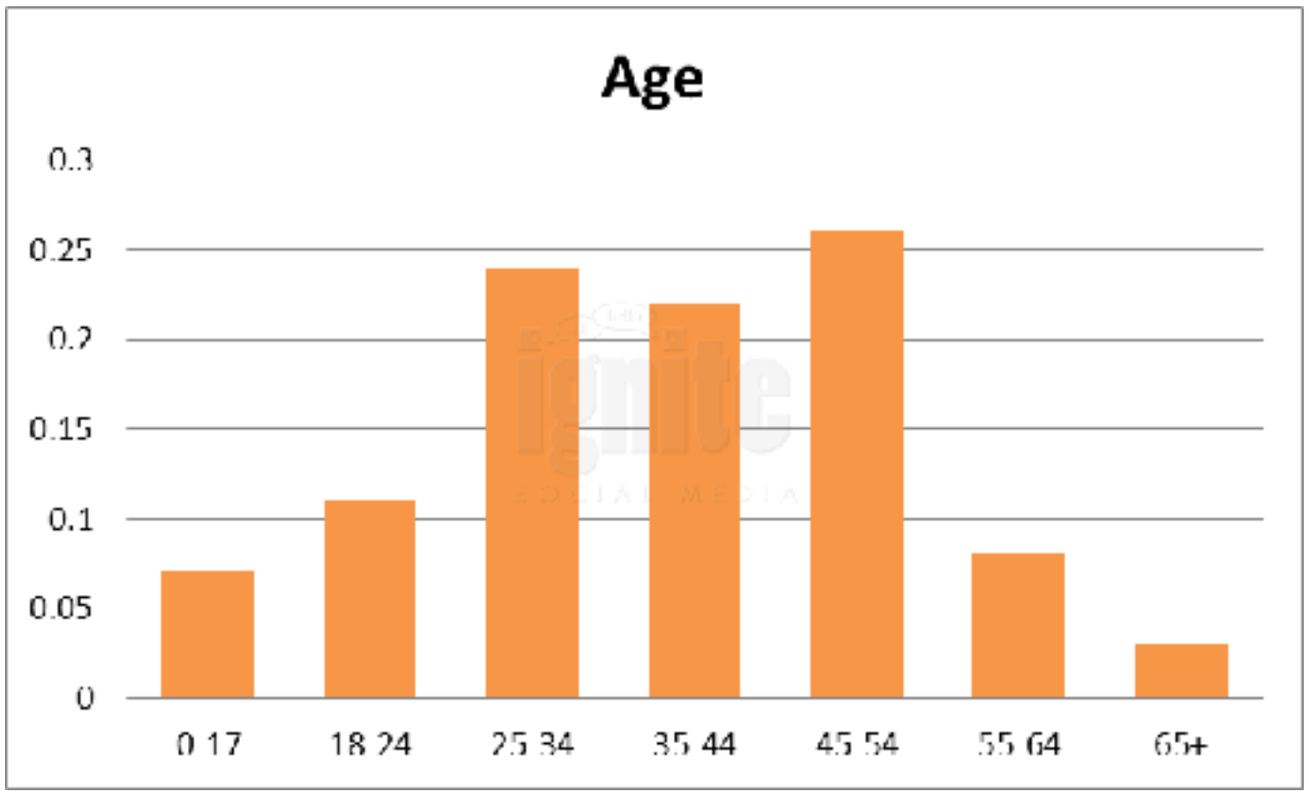
- Quantile-quantile plot (Q-Q plot)



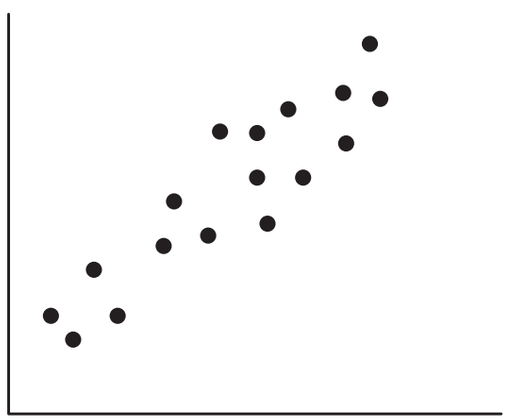
Graphic Displays (2)

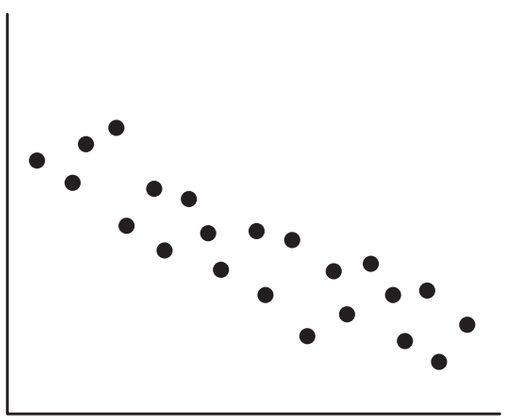
✦ Histogram

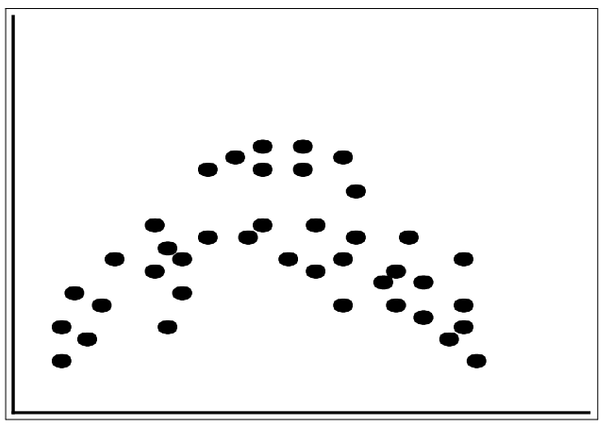
http://www.ignitesocialmedia.com/social-media-stats/2011-social-network-analysis-report/



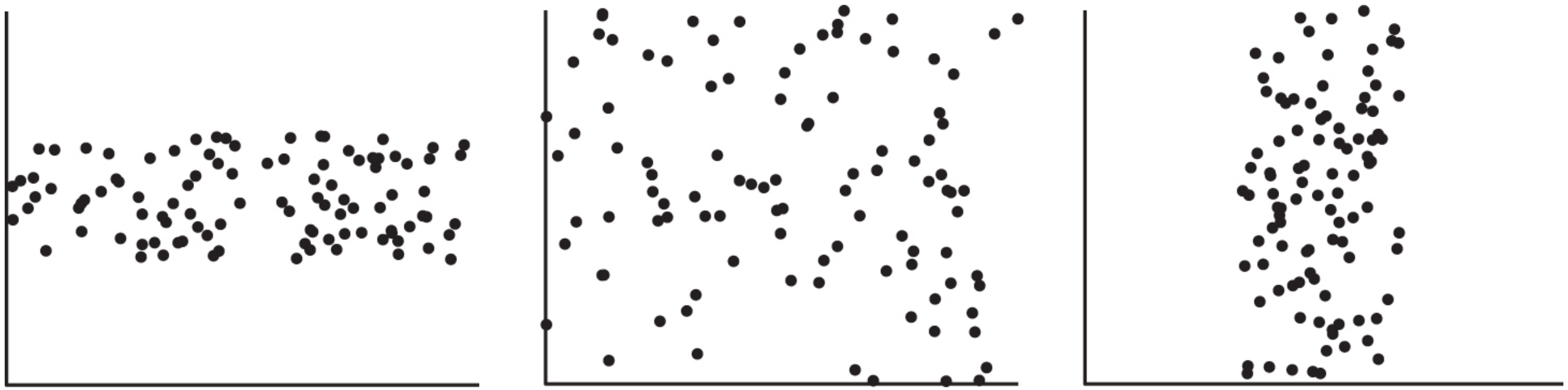
✦ Scatter plot







Not Correlated Data

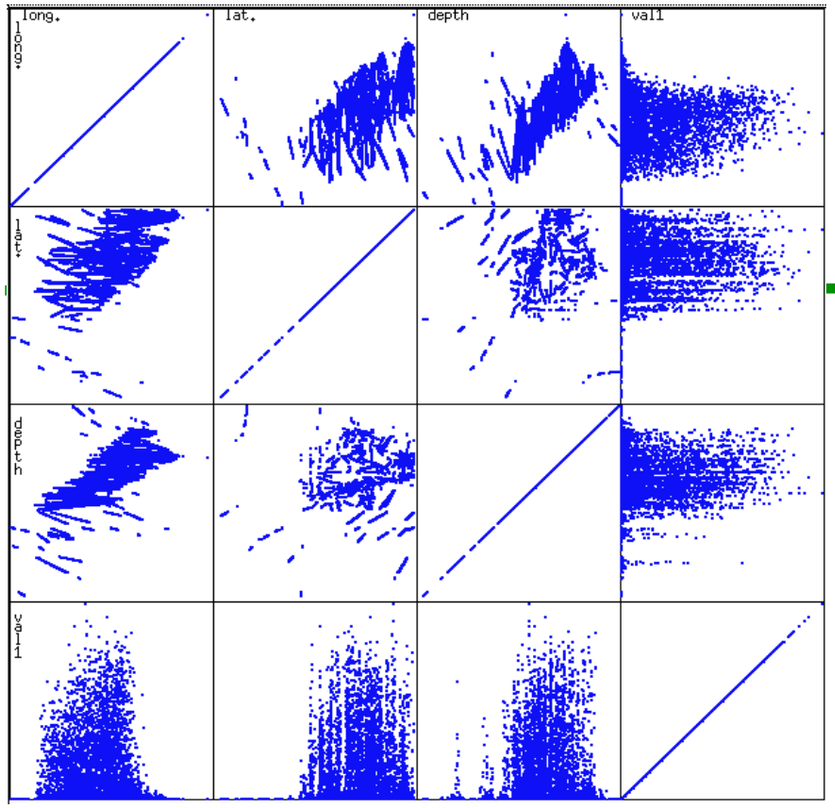


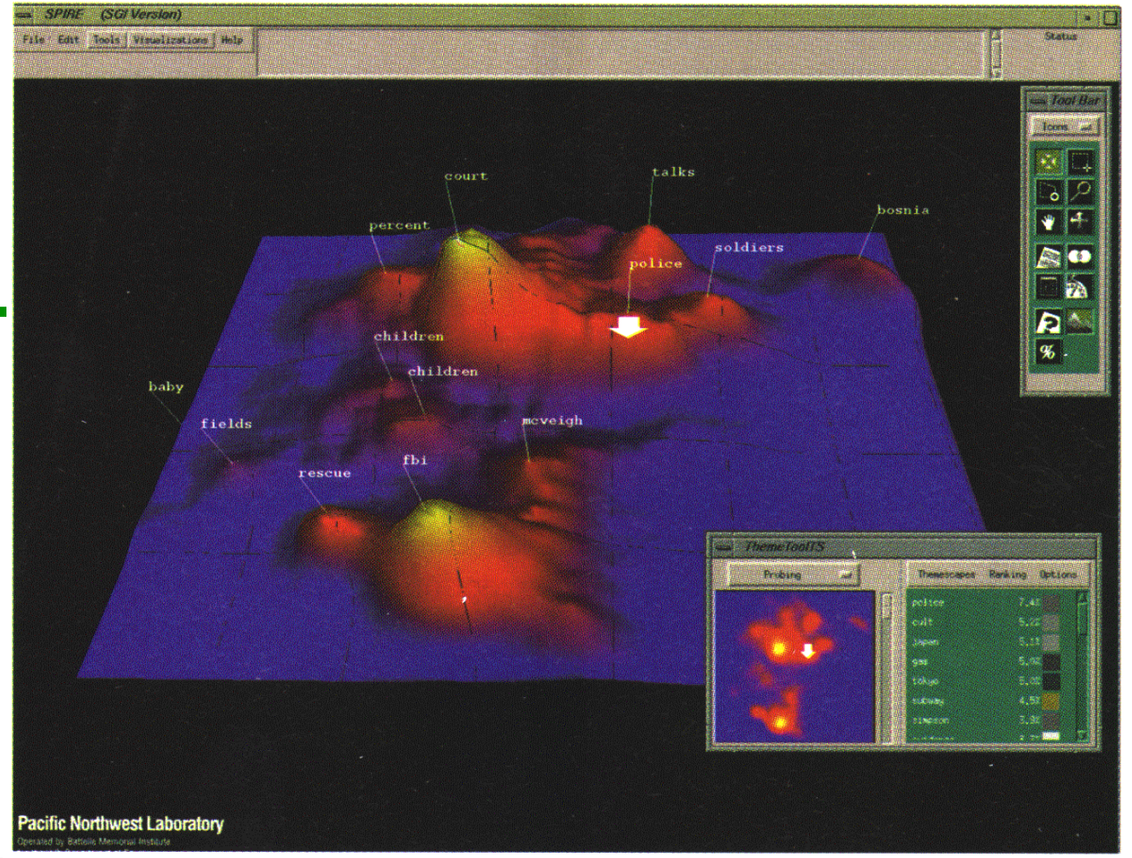
Chapter 2

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

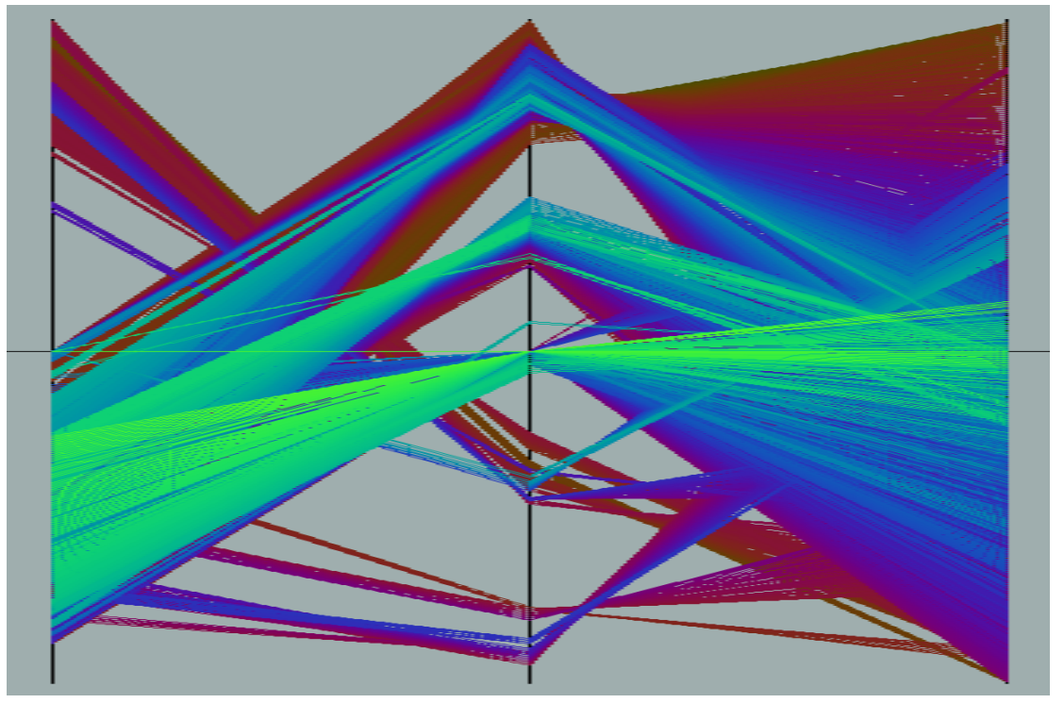
Data Visualization

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


long.
lat.
depth
val1
long.
lat.
depth
val1


SPIRE (SGI Version)
File Edit Tools Visualizations Help
Status
Tool Bar
Icons
court
talks
percent
bosnia
soldiers
police
children
children
baby
fields
mcveigh
rescue
fbi
ThemeTools
Probing
Themescapes Ranking Options
Pacific Northwest Laboratory


How do gun rights backers view gun control after Newtown?
The Supreme Court Ruling on Guns: There Is Something for Both Sides in the ...
Woman stuck between building walls in Portland is pulled to safety
Can 49ers pick up where Seahawks left off?
How Chip Kelly became Eagles new head coach
Marine pleads guilty to urinating on corpses
Obama announces 23 executive actions, asks Congress to pass gun laws
NYC school bus drivers strike over job security
White House ups the ante for petition website 100000 signatures
Oregon sheriff tells Biden he won't enforce new gun laws
Gun rights supporters rallied to state Capitols
UPDATE 2-Massachusetts governor offers new gun laws after Newtown
Queens Park Rangers complete signing of Marseille striker Loic Remy for club ...
Mike McCoy was a hot commodity. Tim Tebow might not be
Hawks suspend Josh Smith for conduct detrimental to the team
Ray Lewis of Ravens: 'No better way to write it up'
A-Rod has surgery on hip, to be out until July
Oil Rises as US Inventories Unexpectedly Decrease
Horse meat found in 'beef' burgers — reportedly as much as one-third
Pakistan government calls demands by protesters 'not possible'
Egypt's Morsi tries to defuse flap over Jews slur
Netanyahu Issues Veiled Barb in Response to Reported Criticism From Obama
Egyptian residential block collapses, leaving at least 23 dead
Russian court turns down Pussy Riot appeal
Emergency room visits tied to energy drinks have more than doubled since 2007 ...
UPDATE 1-German economic calls for ECB to return to core mandate
US Stocks Drop on World Bank Forecasts Amid Earnings
Learn how to maximize Social Security benefits
Conrad Bain, 'Diff'rent Strokes' dad, dies at 89
A Snub by Oscars? Affleck Has an Answer
Shakira, Gerard Pique host virtual baby shower for underprivileged kids
newsmap



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

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

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)

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

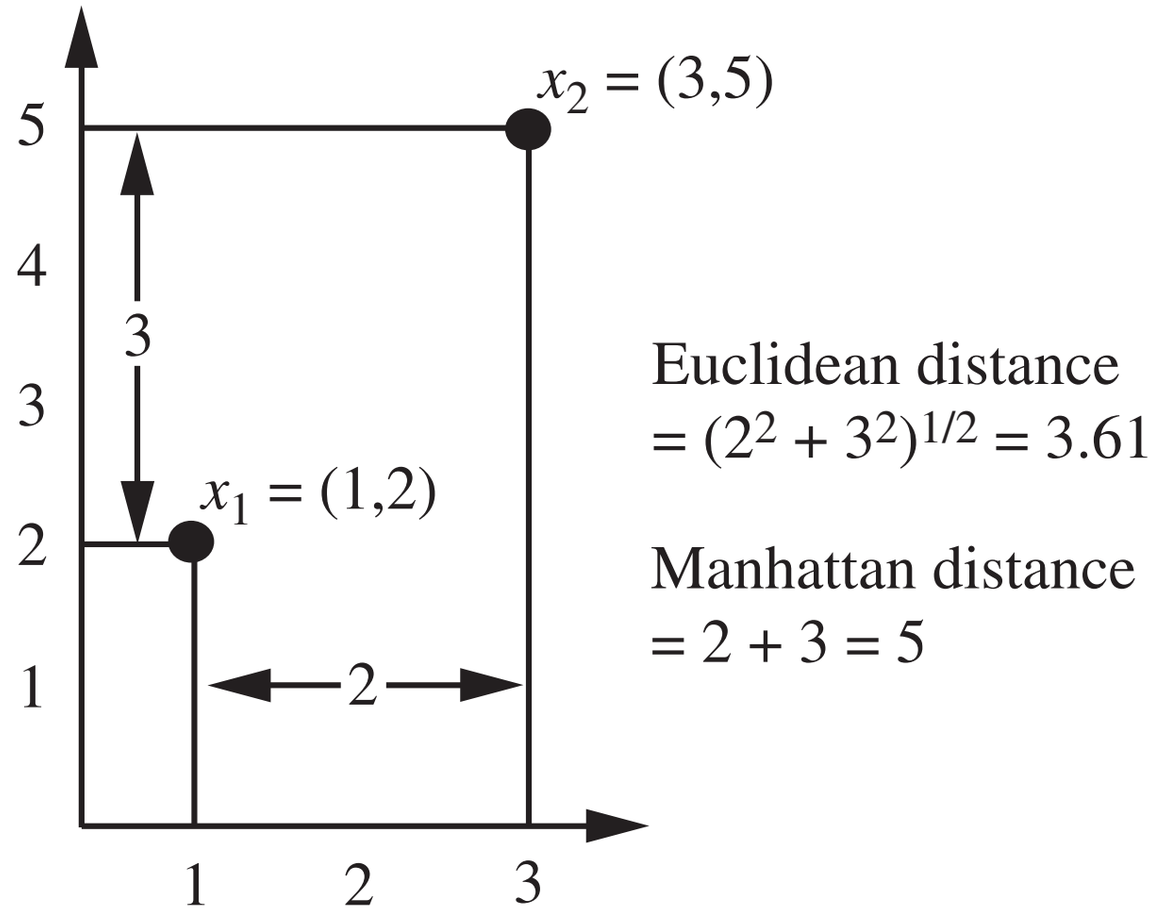
- Manhattan distance (L_1 norm)

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

- 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)

Nominal Attributes

- E.g., hair color, occupation
- Method 1: simple matching
  - d(i, j) = (p - m) / p
  - m: # of matches, p: total # of variables
- Method 2: view each state as a binary variable
  - e.g., colors (red, green, yellow, blue); then (0, 1, 0, 0) means color green

Binary Variables

		obj_j	obj_j	
		1	0	sum
obj_i	1	q	r	q+r
obj_i	0	s	t	s+t
	sum	q+s	r+t	q+r+s+t

- Contingency table

- Distance measure for symmetric binary variables

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Distance measure for asymmetric binary variables

$$d(i,j) = \frac{r+s}{q+r+s}$$

$$sim(i,j) = \frac{q}{q+r+s} = 1 - d(i,j)$$

- Jaccard coefficient

Example

$$d(i,j) = \frac{r+s}{q+r+s}$$

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric binary variable
- Others are asymmetric binary variables
- Consider only asymmetric binary variables
  - Y (yes) and P (positive) is 1, and N is 0
    - d(Jack, Mary) = (0+1)/(2+0+1) = 0.33
    - d(Jack, Jim) = (1+1)/(1+1+1) = 0.67
    - d(Jim, Mary) = (1+2)/(1+1+2)=0.75

Ordinal Variables

- Example: gold, silver, bronze
- Order is important: rank
- Treat like interval-scaled variables
  - map to their ranks $r_{if} \in \{1, \ldots, M_f\}$
  - map to range [0, 1] $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
  - e.g., (1, 2, 3) => (0.0, 0.5, 1.0)
  - dissimilarity of interval-scaled variables

Variables of Mixed Types

- Data may contain different types of variables
  - interval-scaled, binary (symmetric or asymmetric), categorical, ordinal, ratio
  - Weighted combination
    - $\delta_{ij}^{(f)} = 0$ if
      - x_{if} or x_{jf} is missing
      - $x_{if} = x_{jf} = 0$ and f is asymmetric binary
    - otherwise = 1

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

Cosine Similarity Example

- D1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
- D2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$$s(x, y) = \frac{x^t \cdot y}{||x|| ||y||}$$

- D1 • D2 = 5x3 + 0x0 + … + 0x1 = 25
- ||D1|| = (5x5 + 0x0 + … + 0x0)^(1/2) = 6.481
- ||D2|| = 4.12, cos(D1, D2) = 0.936

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Summary

- Chapter 2: Getting to know your data
  - data objects and attribute types
  - basic statistical description of data
  - data visualization
  - measuring data similarity and dissimilarity

Announcements

- Office hours
  - Tu 11am to 12pm; Fr 1pm to 2pm
  - ECCR 1B24, zoom, or by appointment
  - TA: Tu 4pm to 5pm, ECCR 1B10
- Homework 1
  - will be posted at moodle on Thursday
  - due at 9:30am, Thursday Sep 12
  - submit at moodle