Data Mining:

Concepts and Techniques

— Chapter 2 —

Jiawei Han, Micheline Kamber, and Jian Pei

University of Illinois at Urbana-Champaign

Simon Fraser University

Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

	team	coach	pla y	ball	score	game	wi n	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

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

- One should know about data(attributes and values)
- Fixing inconsistencies in data integration
- Easy to fill in missing values
- Easy to smooth noisy values
- Spot outliers
- Know whether data is symmetric or skewed i.e distribution and dispersion of data

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples* , *examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

Attributes

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

- **Nominal:** categories, states, or "names of things", do not have any meaningful order, enumeration
  - *Hair_color = {auburn, black, blond, brown, grey, red, white}*
  - marital status, occupation
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - <u>Symmetric binary</u>: both outcomes equally important, have same weight
    - e.g., gender
  - <u>Asymmetric binary</u>: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size = {small, medium, large}*, grades{A,B,C,D}, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
  - Can compare and quantify
  - Numeric in nature
  - Measures of central tendency(mean, median, mode)
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

- **Attributes ML point of view**
- **Discrete Attribute**
  - Has only a finite or countably infinite(one to one correspondence with natural number)set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

  Note: *n* is sample size and *N* is population size.

  - Weighted arithmetic mean:

  $$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Trimmed mean: chopping extreme values
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

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

  - L1- lower bound of median interval, n- number of values in entire dataset, $(\sum freq)_l$ is sum of frequencies of all intervals lower than median interval, $freq_{median}$ is median interval frequency and width is width of median interval

age	frequency
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Median interval → 21–50

- **Mode**
  - Mode for data set is a value that occurs most frequently.
  - Unimodal, bimodal, trimodal
  - Empirical formula: for unimodal numeric data that is moderately skewed.

$$mean - mode = 3 \times (mean - median)$$

With mean and mode values known, we can approximate mode for skewed data.

- **Midrange**

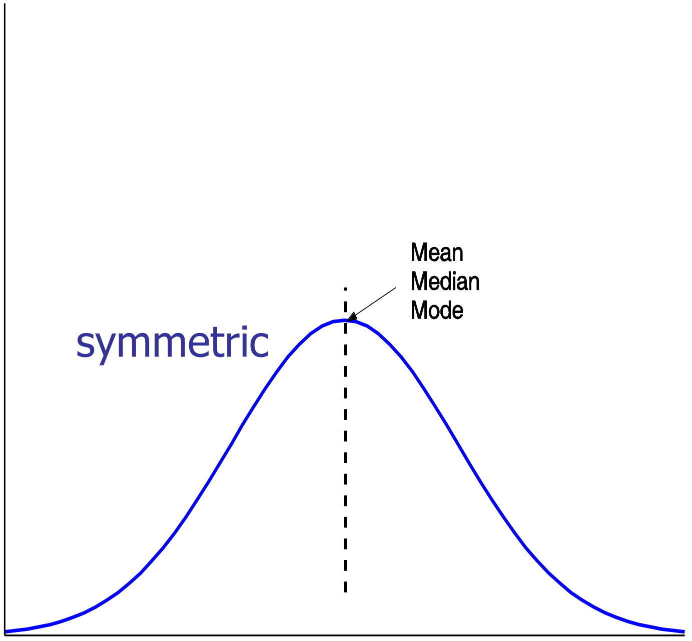
It is the average of the largest and smallest values in the set. This measure is easy to compute using the SQL aggregate functions, max() and min().

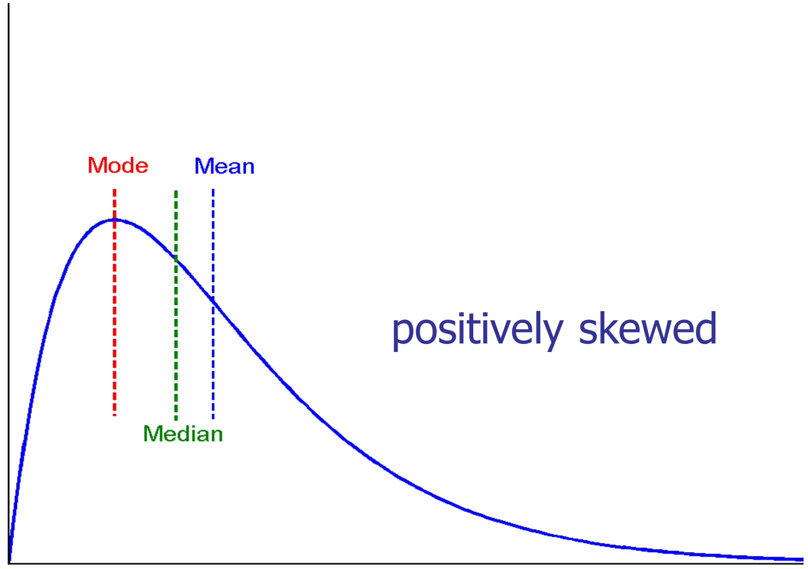
E.g. values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

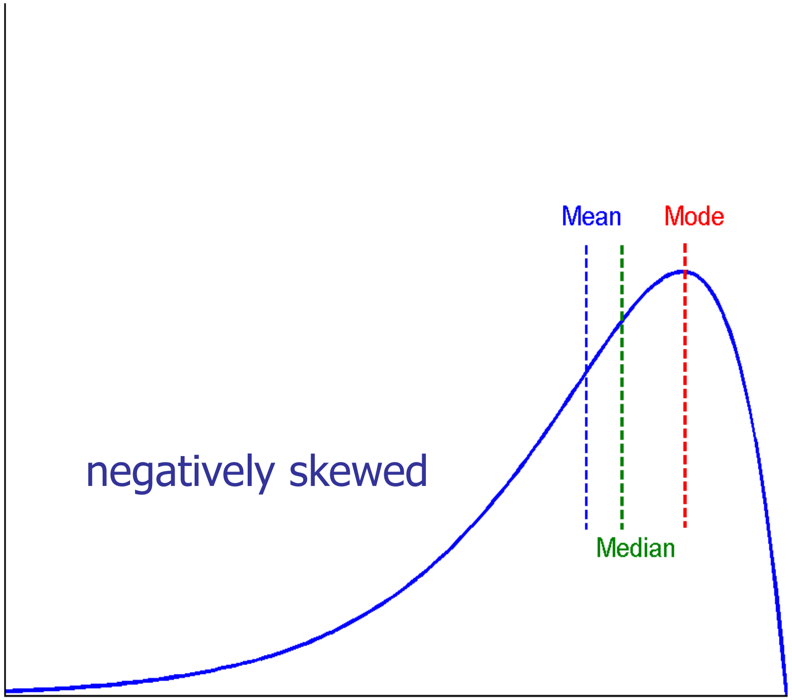
midrange of the data of above example is (30,000+110,000)/2 = $70,000.

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed unimodal data
- "In a skewed distribution..., the mean is pulled in the direction of the extreme scores or tail (same as the direction of the skew), and the median is between the mean and the mode."



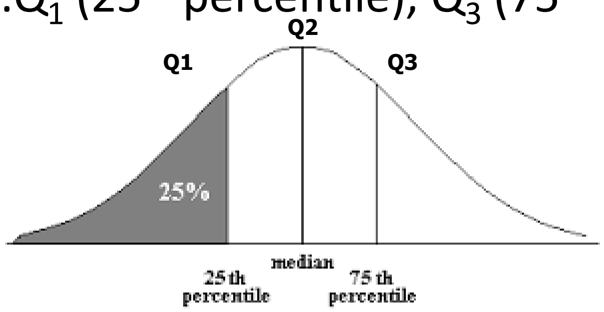




Measuring the Dispersion of Data

- Quantile, Quartiles, outliers and boxplots
  - **Quantiles** are data points taken at regular intervals of data distribution
  - **Quartiles**: quantiles i.e 3 data points dividing data distribution in four equal parts

E.g.Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)



  - **Inter-quartile range**: IQR = $Q_3 - Q_1$

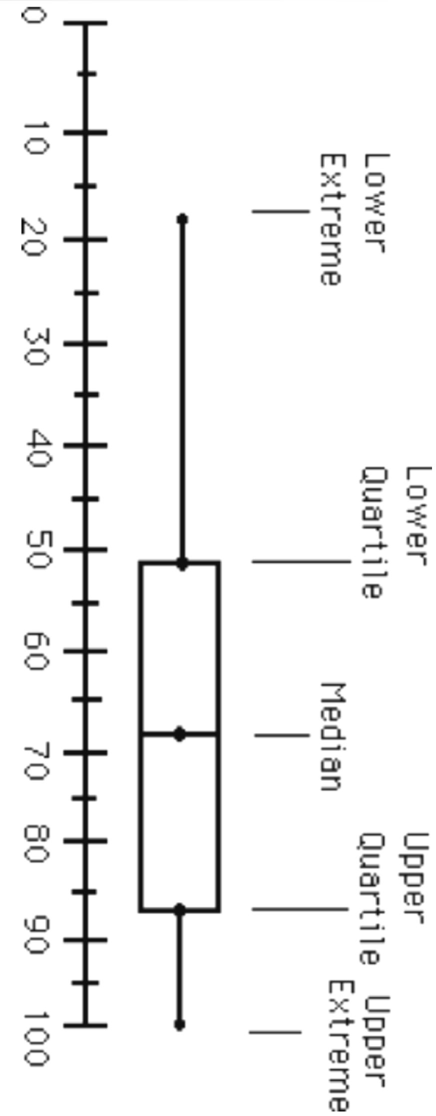
Simple measure of spread that gives range covered by middle half of the data

Measuring the Dispersion of Data

- **Five number summary**:in symmetric distribution median and other median splits the data into equal size halves.
- Not true with skewed distributions
- As Q_1, median, Q_3 does not give idea about end points of data.
- min, Q_1, median, Q_3 max
- **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
- **Outlier**: usually, a value higher/lower(above 3rd /below first quartile) than 1.5 x IQR

Boxplot Analysis

- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually



Outliers

Mild vs. Extreme Outliers
Extreme outliers are data points that are more extreme than Q1 - 3 * IQR or Q3 + 3 * IQR.

Extreme outliers are marked with an asterisk (*) on the boxplot.

Mild outliers are data points that are more extreme than than Q1 - 1.5 * IQR or Q3 + 1.5 * IQR, but are not extreme outliers.

Mild outliers are marked with a circle (O) on the boxplot.

Example of Boxplot

- Compute Q1, Q2 and Q3. Also, compute the interquartile range IQR = Q3 - Q1.

Example: Suppose that the dataset consists of these hypothetical test scores:

5 39 75 79 85 90 91 93 93 98

Q1 = 75, Q2 = 88, Q3 = 93. IQR = 93 - 75 = 18.

- Draw three horizontal lines, all of the same length and all starting at the same x-value: one at height Q1, the second at Q2 (median) and the third at Q3.
- Draw two vertical lines, one at connecting the left endpoints of the lines and the other connecting their right endpoints.
- Compute the inner fences IF1 = Q1 - 1.5 * IQR and IF2 = Q3 + 1.5 * IQR.

Example: The inner fences are

IF1 = 75 - 1.5 * 18 = 48 and IF2 = 93 + 1.5 * 18 = 120.

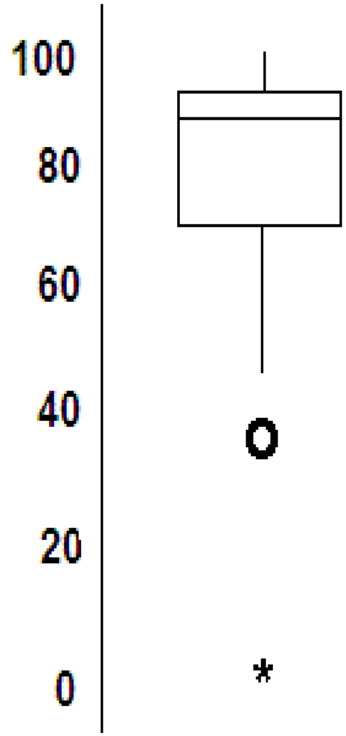
- Draw a whisker downward from Q1 to IF1 or Q0, whichever comes first. Draw a whisker upward from Q3 to IF2 or Q4, whichever comes.
- Compute the outer fences OF1 = Q1 - 3 * IQR and OF2 = Q3 + 3 * IQR.

Example: The outer fences are

OF1 = 75 - 3 * 18 = 21 and OF2 = 93 + 3 * 18 = 147.

- Extreme outliers are observations that are beyond one of the outer fences OF1 or OF2. Mark any extreme outliers on the boxplot with an asterisk (*).**Example:** The only observation less than OF1 = 21 is 5.
- Mild outliers are observations that are between an inner and outer fence. Mild outliers are marked with a circle (O).**Example:** The only observation that is between an inner fence and an outer fence is 39, which is between IF1 = 48 and OF1 = 21. O.

- Box plot example



Data set:
5 39 75 79 85 90 91 93 93 98

Q1 = 75, Q2 = 88, Q3 = 93. IQR = 93 - 75 = 18.

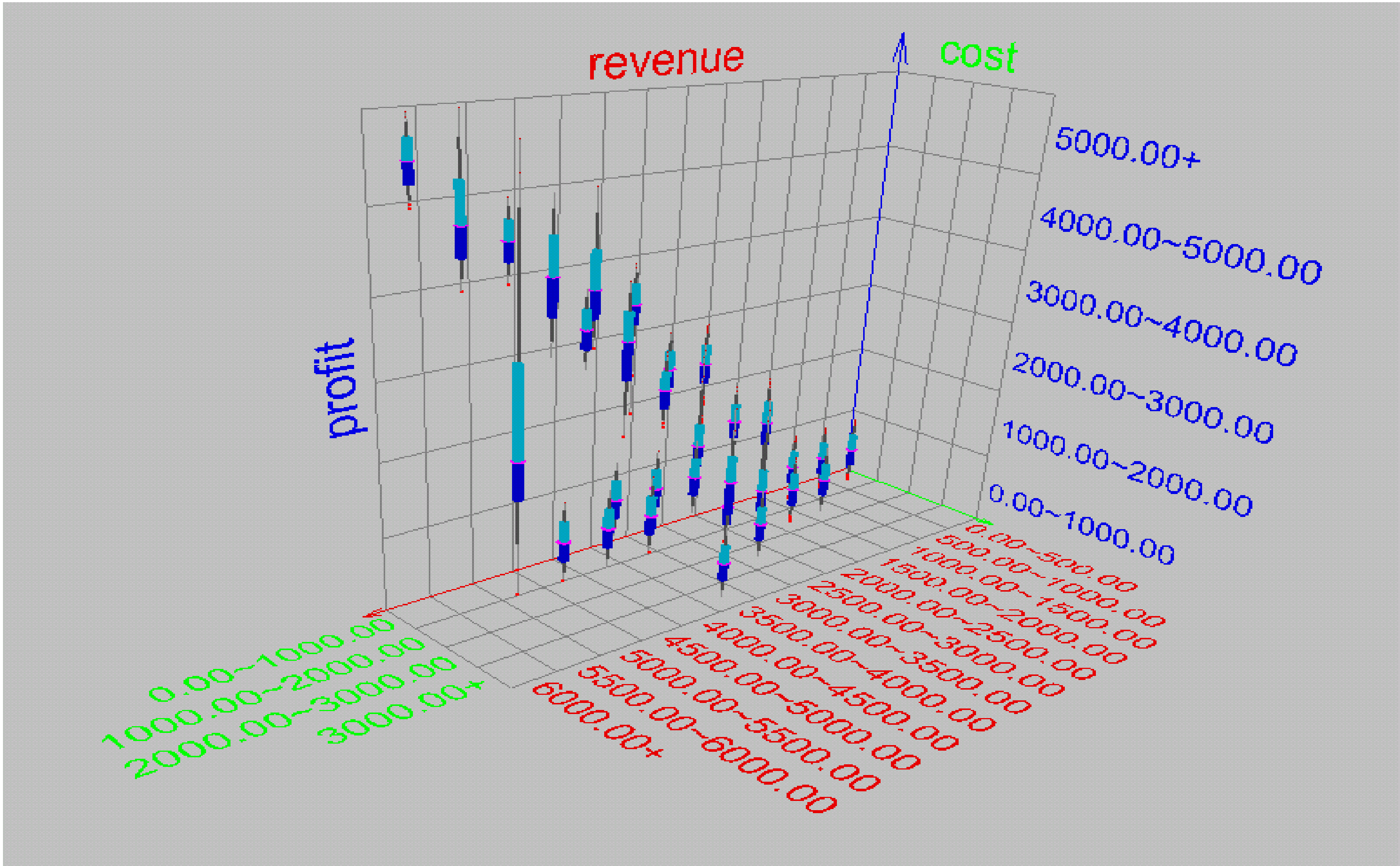
IF1 = 75 - 1.5 * 18 = 48 and IF2 = 93 + 1.5 * 18 = 120.

OF1 = 75 - 3 * 18 = 21 and OF2 = 93 + 3 * 18 = 147.

122 IN PLACE OF 98?

- Suppose that the data for analysis includes the attribute age. The age values for the data
- tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30,33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
- (a) What is the mean of the data?What is the median?
- (b) What is the mode of the data? Comment on the data's modality (i.e., bimodal,
- trimodal, etc.).
- (c) What is the midrange of the data?
- (d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
- (e) Give the five-number summary of the data.
- (f) Show a boxplot of the data.

Visualization of Data Dispersion: 3-D Boxplots



Variance and standard deviation

- Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is. A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that

  the data are spread out over a large range of values
- The **variance** of *N* observations, $x1, x2, : : : , xN$, for a numeric attribute *X* is

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

  - Standard deviation σ is the square root of variance σ^2

E.g.

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12}$$

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is \$58,000.

$$\sigma^2 = \frac{1}{12}(30^2 + 36^2 + 47^2 \ldots + 110^2) - 58^2$$

$$\approx 379.17$$

$$\sigma \approx \sqrt{379.17} \approx 19.47.$$

2·28. Scores of two golfers for 24 rounds were as follows :

Golfer A : 74, 75, 78, 72, 77, 79, 78, 81, 76, 72, 72, 77, 74, 70, 78, 79, 80, 81, 74, 80, 75, 71, 73.

Golfer B :86, 84, 80, 88, 89, 85, 86, 82, 82, 79, 86, 80, 82, 76, 86, 89, 87, 83, 80, 88, 86, 81, 81, 87.

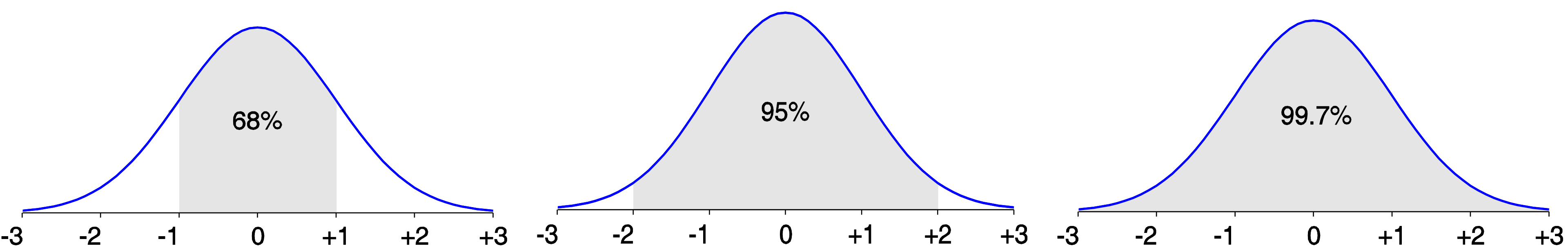
Find which golfer may be considered to be more consistent player ?

2·29. The sum and su...

- The basic properties of the standard deviation, σ, as a measure of spread are as follows:
- σ measures spread about the mean and should be considered only when the mean is chosen as the measure of center.
- $\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise, $\sigma > 0$.

Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From $\mu-\sigma$ to $\mu+\sigma$: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From $\mu-2\sigma$ to $\mu+2\sigma$: contains about 95% of it
  - From $\mu-3\sigma$ to $\mu+3\sigma$: contains about 99.7% of it

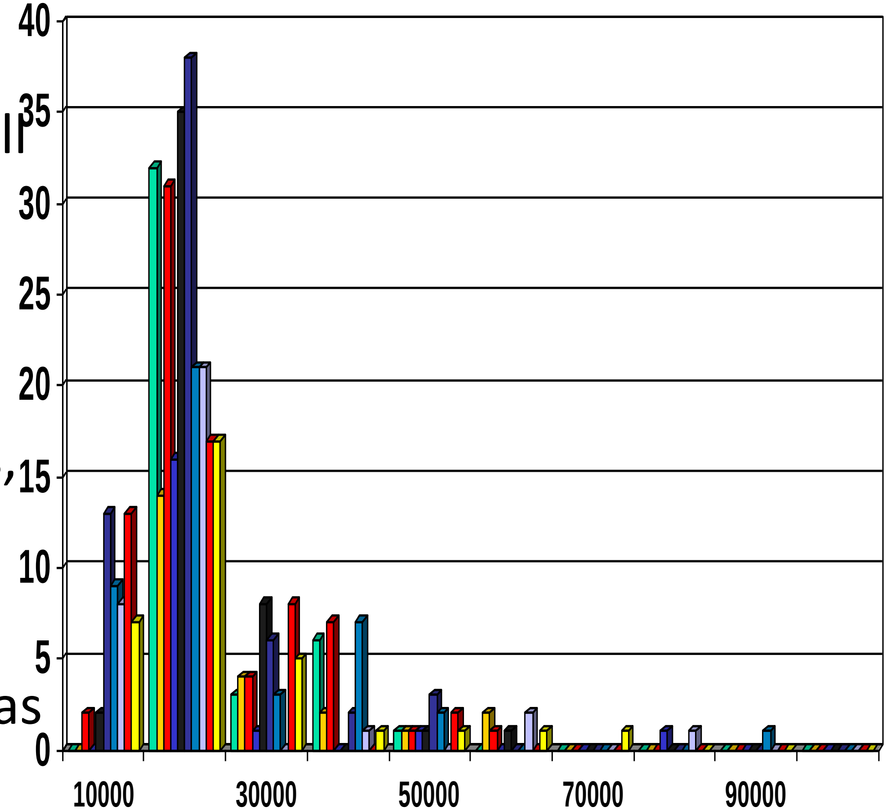


Graphic Displays of Basic Statistical Descriptions

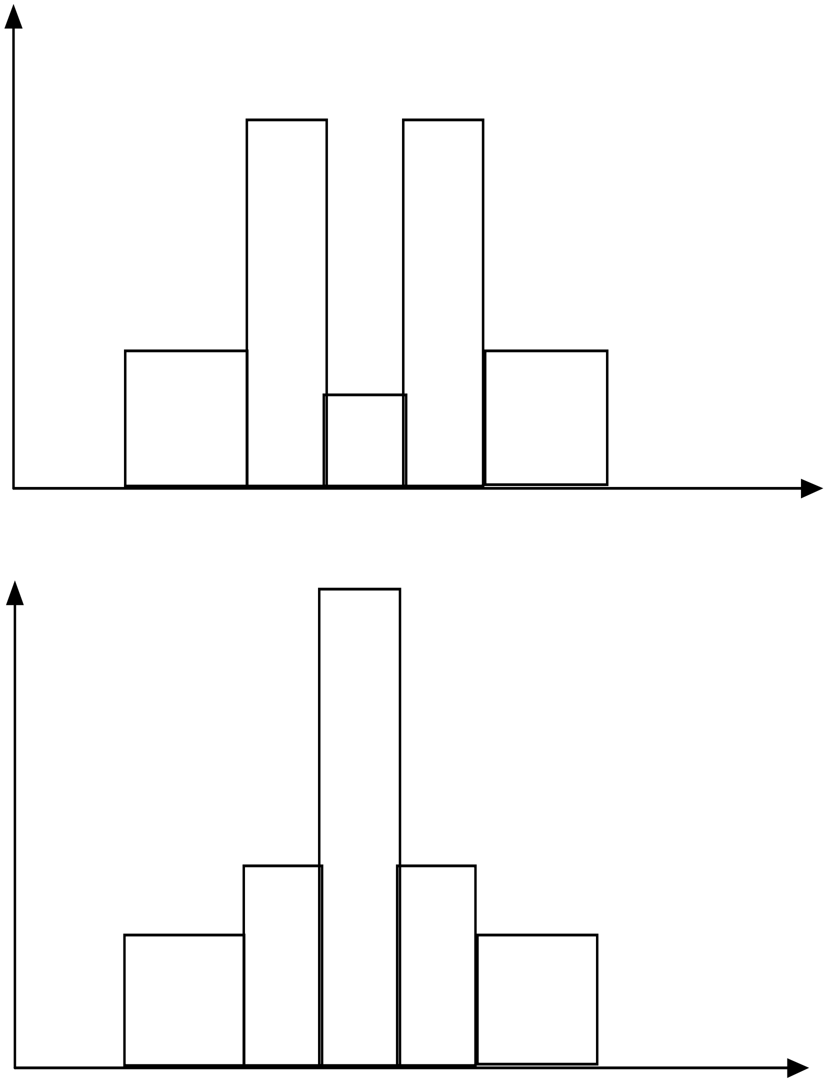
- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately $100\, f_i\,\%$ of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



Histograms Often Tell More than Boxplots

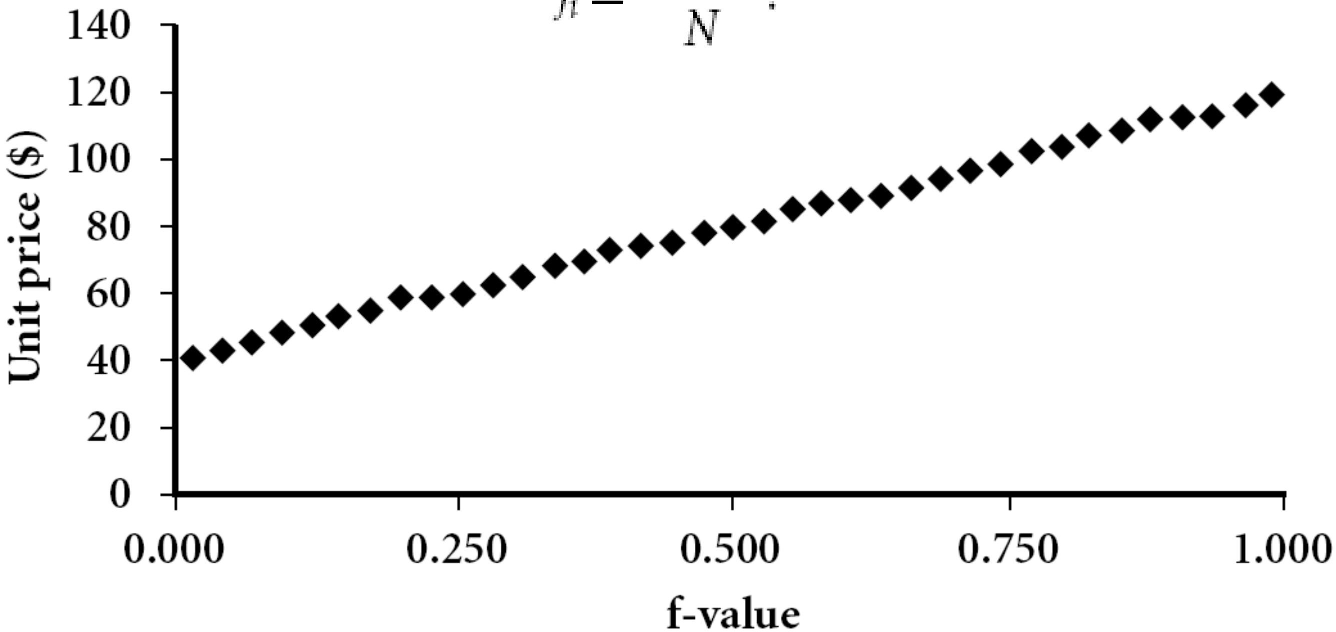


- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

Quantile Plot

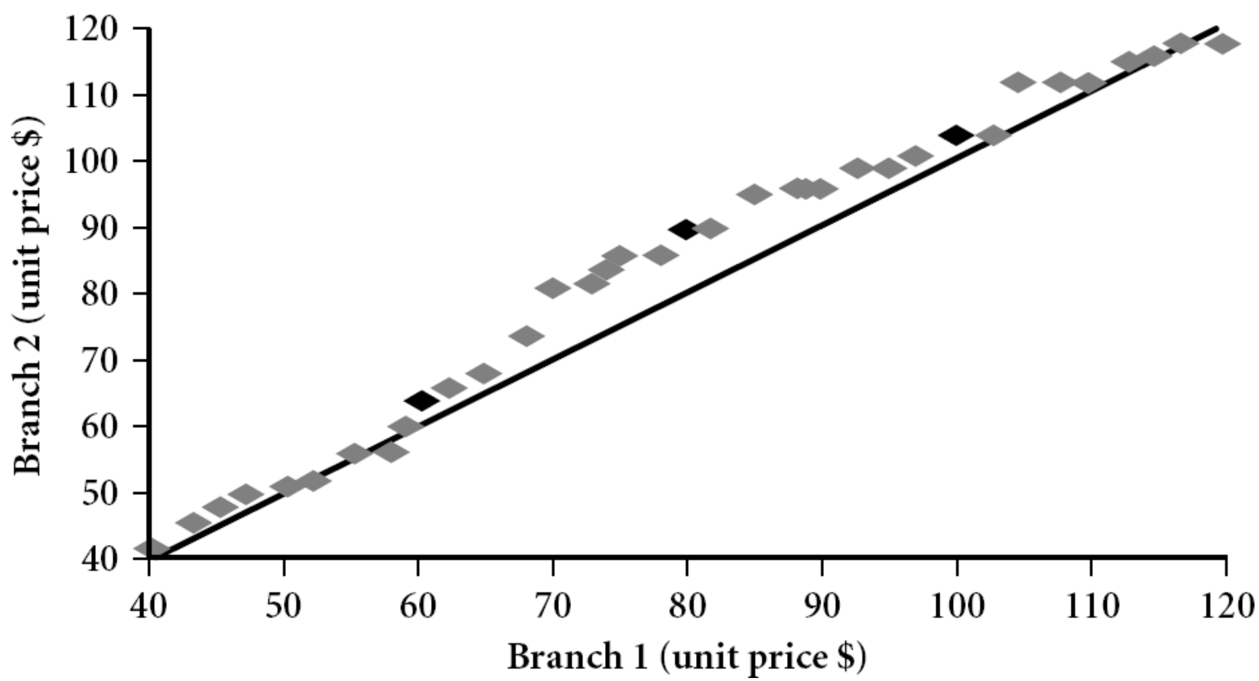
- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i

$$f_i = \frac{i - 0.5}{N}.$$



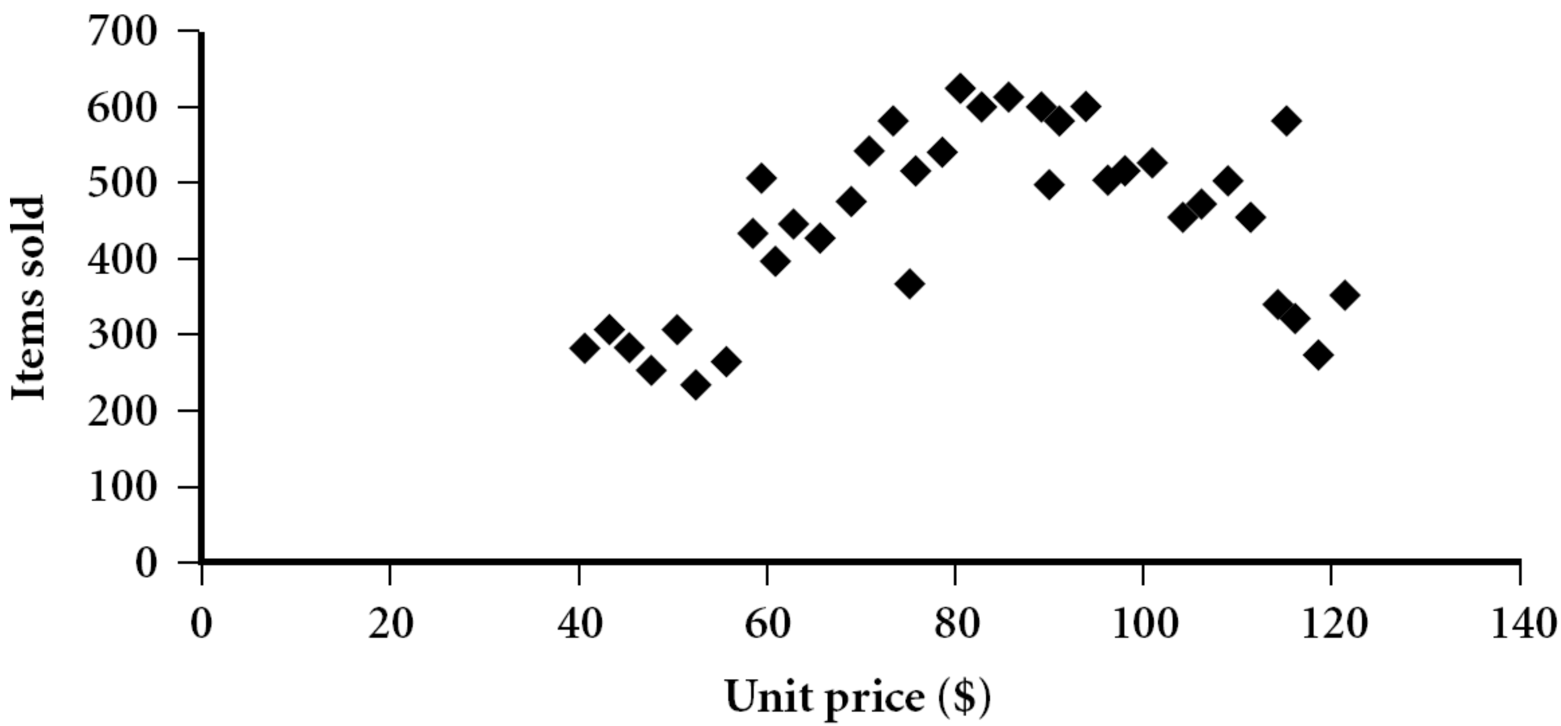
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

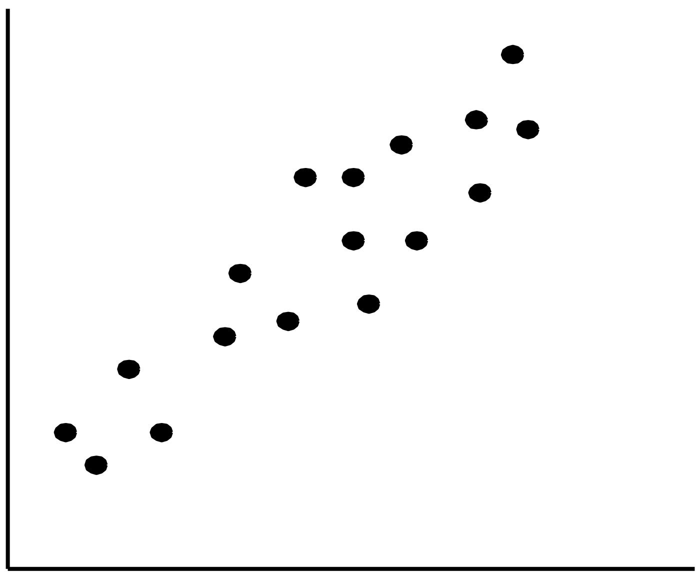


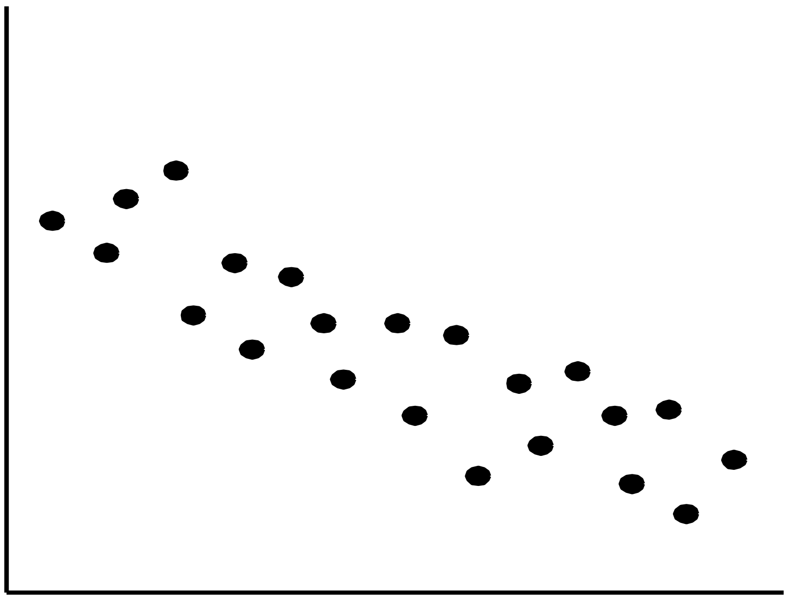
Scatter plot

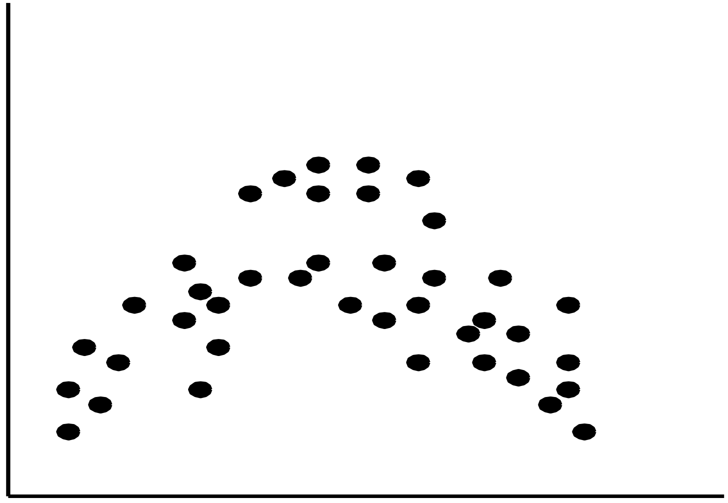
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data

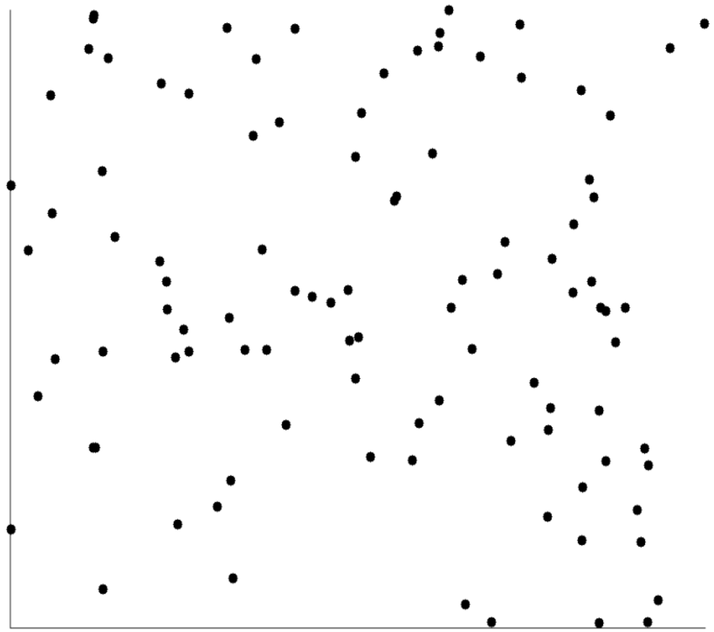


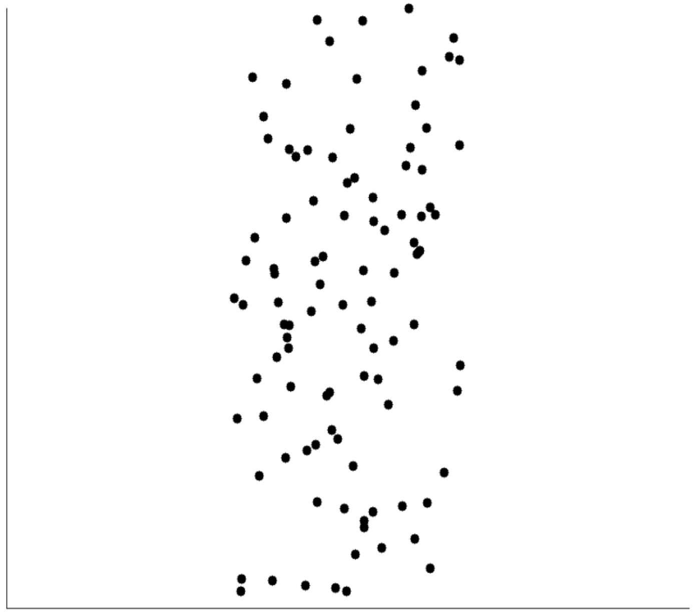




- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data





Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary