Data Mining:

Concepts and Techniques

— Chapter 2 —

Jiawei Han, Micheline Kamber, and Jian Pei
University of Illinois at Urbana-Champaign
Simon Fraser University

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**

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.
- **Alternate names:** dimension, feature and variable
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Ordinal
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types (set of possible values)

1. Nominal: symbols,categories, states, or “names of things”

- *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
- marital status, occupation, ID numbers, zip codes
- Don’t have any meaningful order

2. Binary

- Nominal attribute with only 2 states (0 and 1)
- Symmetric binary: both outcomes equally important
  - e.g., gender
- Asymmetric binary: outcomes not equally important.
  - e.g., medical test (positive vs. negative)
  - Convention: assign 1 to most important outcome (e.g., HIV positive)

3. Ordinal

- Values have a meaningful order (ranking) but magnitude between successive values is not known.
- *Size* = {*small, medium, large*}, grades, army rankings

Attribute Types (set of possible values)

4. Numeric Attribute:

- Quantity (integer or real-valued)

- **Interval-scaled**
  - Measured on a scale of **equal-sized units**
  - Values have order and can be +ve,0,-ve
    - E.g., *temperature in C°or F°, calendar dates*
  - No true zero-point
- **Ratio-scaled**
  - 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

- **Discrete Attribute**
  - Has only a finite or countably infinite 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

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread

- Data dispersion characteristics
  - mean, median, max, min, quantiles, outliers, variance, etc.

Measuring the Central Tendency

- Mean (algebraic measure):

  Note: *N* is number of observations.

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

  - Weighted arithmetic mean:

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

  - Even a small number of extreme values can corrupt the mean.
  - Trimmed mean: chopping extreme values

Q. Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. What is mean value?

Measuring the Central Tendency

- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - **Q. Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. What is median value?**
  - Estimated by interpolation (for *grouped data*):

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

- L1 = the lower boundary of the median interval
- n = number of vales in entire dataset
- $\sum(freq)_1 = sum\ of\ the\ frequencies\ of\ all\ of\ the\ intervals\ that\ are\ lower\ than\ the\ median\ interva$
- $(\text{freq})_{\text{median}}$ = frequency of the median interval

Measuring the Central Tendency

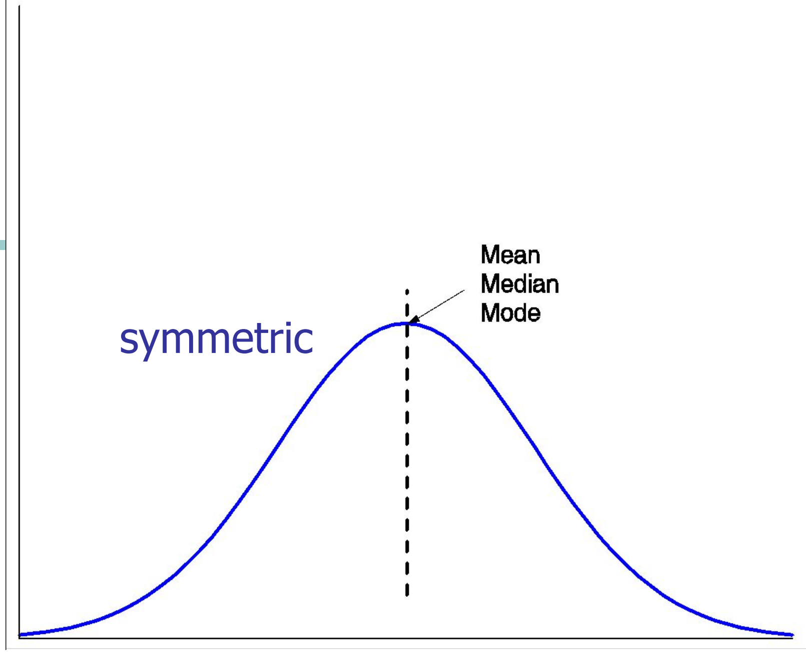
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

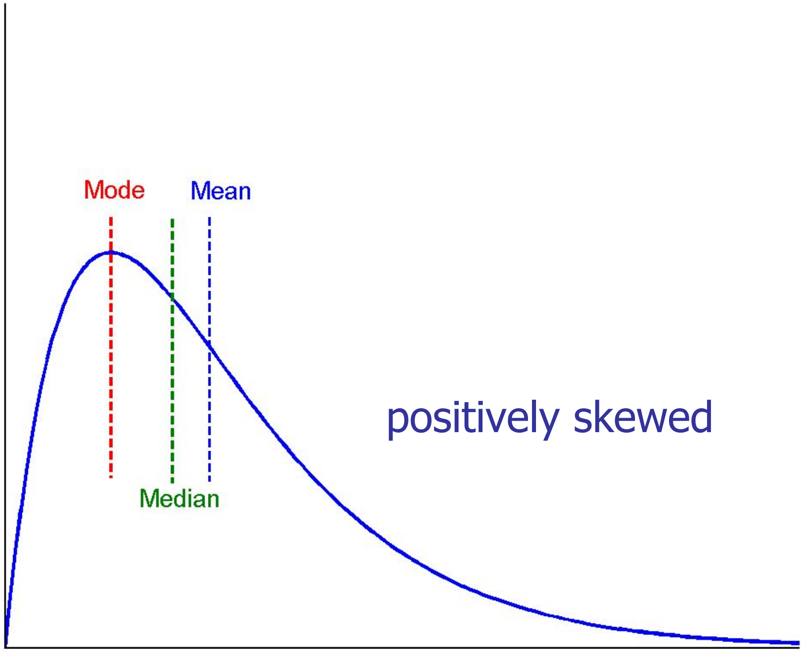
$$mean - mode = 3 \times (mean - median)$$

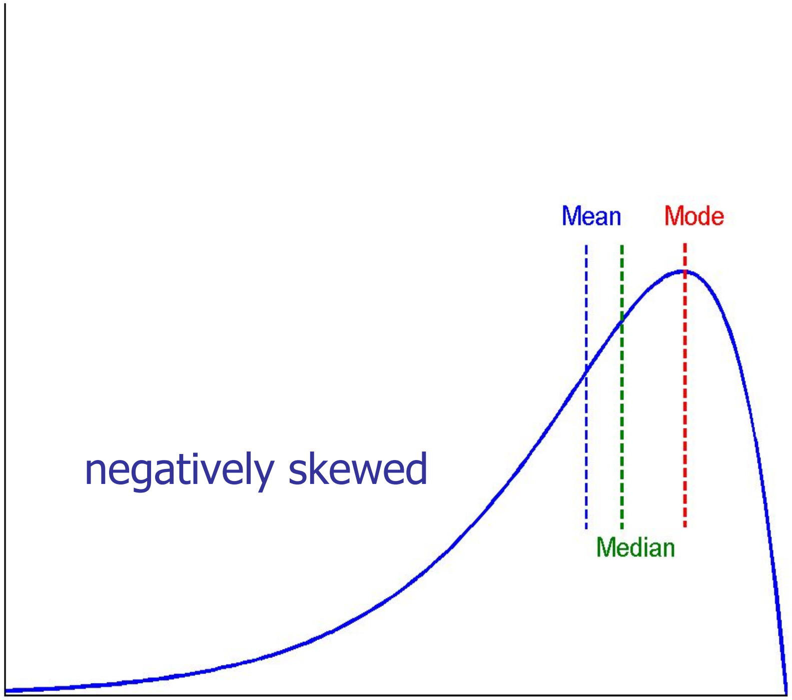
- Midrange:
  - avg of largest and smallest value

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







Variance – ungrouped data

- Variance and standard deviation (*sample: s, population: σ*)
  - **Variance**: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 (or σ^2)

Variance of ungrouped data

- Find the variance for the following set of data representing trees heights in feet:
- **3, 21, 98, 203, 17, 9**

Variance – grouped data

- Variance and standard deviation (*sample: s, population: σ*)
  - **Variance**: (algebraic, scalable computation)

$$s^2 = \frac{\sum_{i=1}^{n} f(m_i - \bar{x})^2}{N-1}$$

  - f = frequency of the class
  - m = midpoint of the class

$$\sigma^2 = \sum f\,(m - \bar{x})^2 / n$$

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 ($or\ \sigma^2$)

Example of Variance (Grouped data)

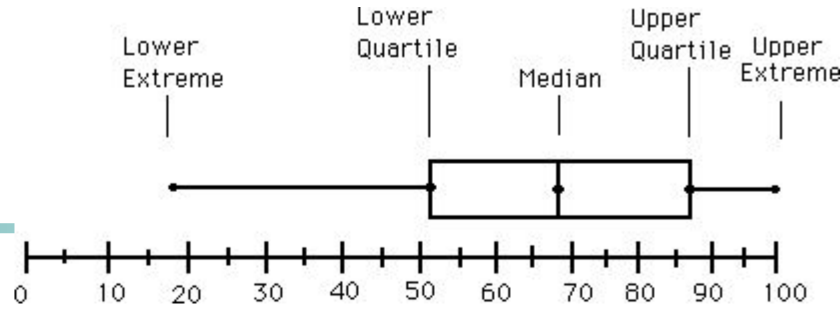
Range	Frequency
1-10	2
11-20	7
21-30	10
31-40	3
41-50	1

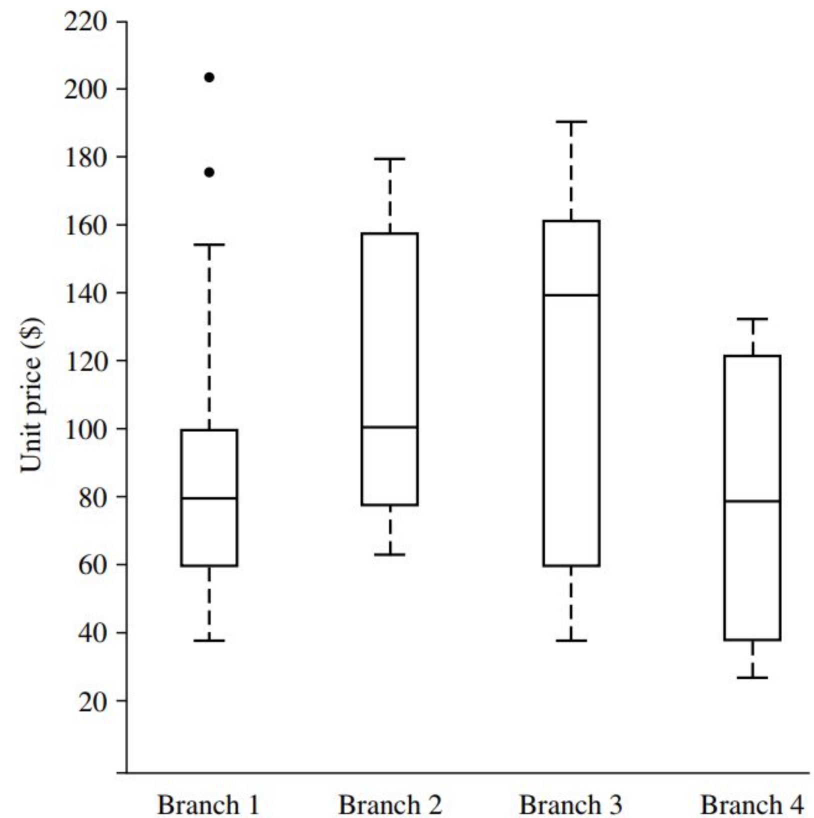
Example - Solution

Range	Frequency (n_i)	Midpoint (m_i)	$m_i * n_i$	μ	$m_i - \mu$	$(m_i - \mu)^2$	$n_i(m_i - \mu)^2$
1-10	2	5.5	11	22.89	-17.39	302.41	604.82
11-20	7	15.5	108.5	22.89	-7.39	54.61	382.28
21-30	10	25.5	255	22.89	2.61	6.81	68.12
31-40	3	35.5	106.5	22.89	12.61	159.01	477.04
41-50	1	45.5	45.5	22.89	22.61	511.21	511.21

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





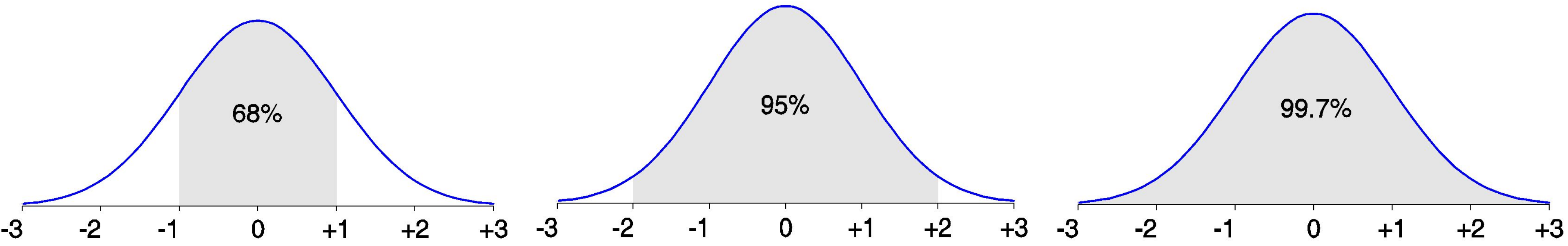
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: 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 than 1.5 x IQR

- **Q. Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.**
- **What is Q1,Q3?**
- **What is IQR?**
- **What is Standard dev?**

Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: 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
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Histograms

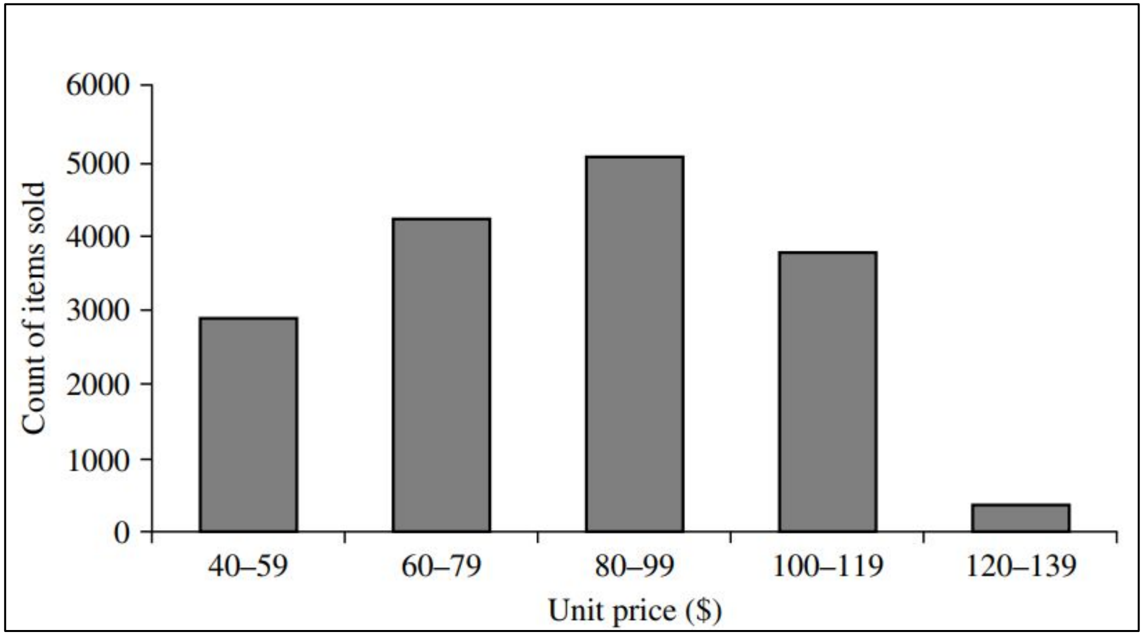
- Graphical method for summarizing the distribution
- If X is nominal, such as **automobile_model** or **item type**, then a pole or vertical bar is drawn for each known value of X.
- The height of the bar indicates the frequency (i.e., count) of that X value.
- The resulting graph is more commonly known as a bar chart.

Histograms - Barchart

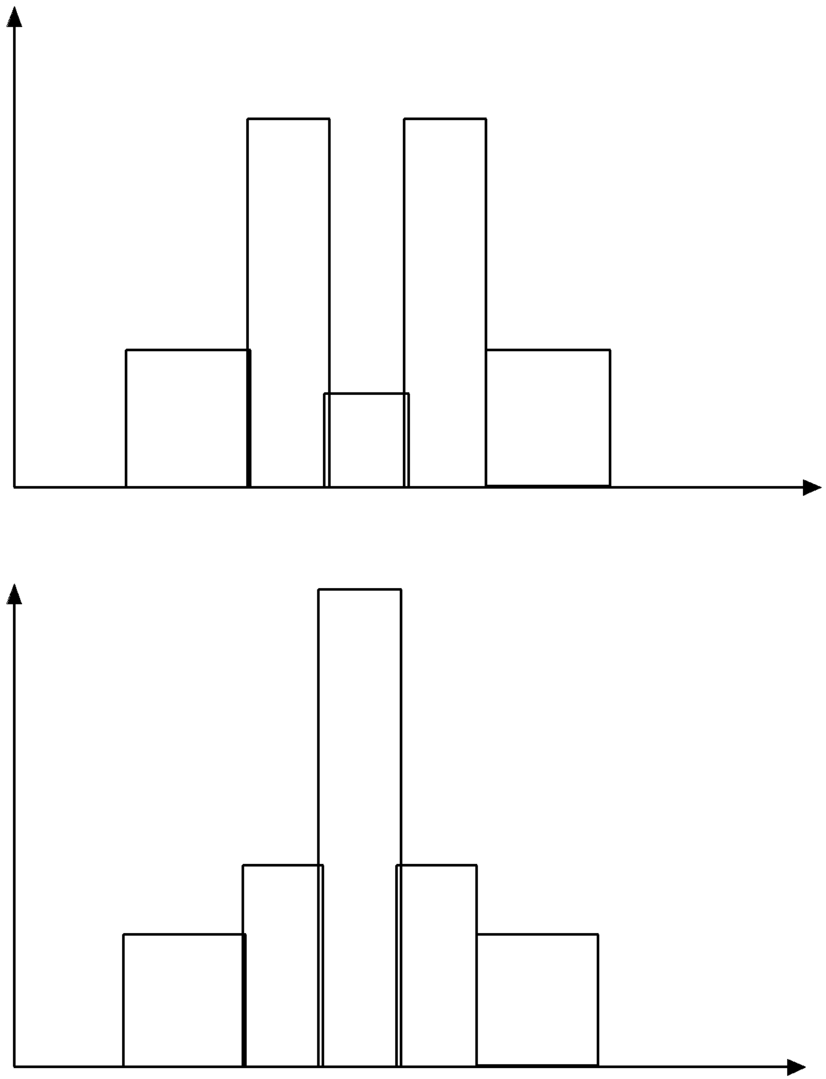
- If X is **numeric**, the term histogram is preferred.
- The range of values for X is partitioned into disjoint consecutive subranges.
- The subranges, referred to as buckets or bins, are disjoint subsets of the data distribution for X.
- The range of a bucket is known as the width. Typically, the buckets are of equal width.
- For example, a price attribute with a value range of $1 to $200 (rounded up to the nearest dollar) can be partitioned into subranges 1 to 20, 21 to 40, 41 to 60, and so on

Histogram

Unit price ($)	Count of items sold
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350



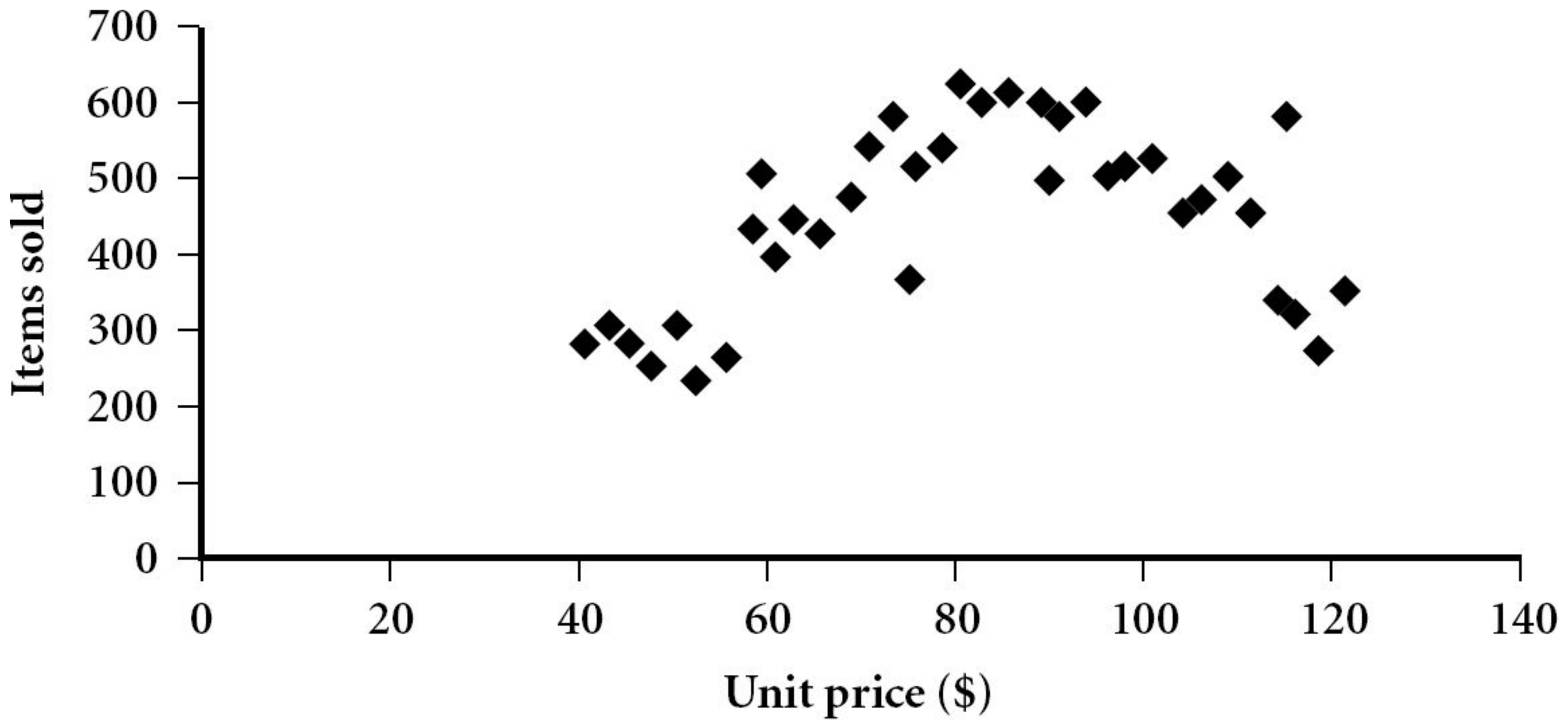
Histograms Often Tell More than Boxplots



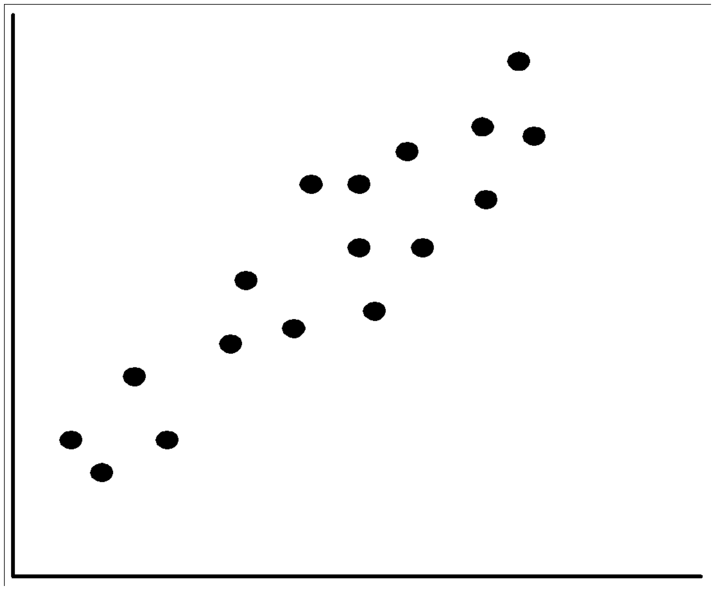
- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

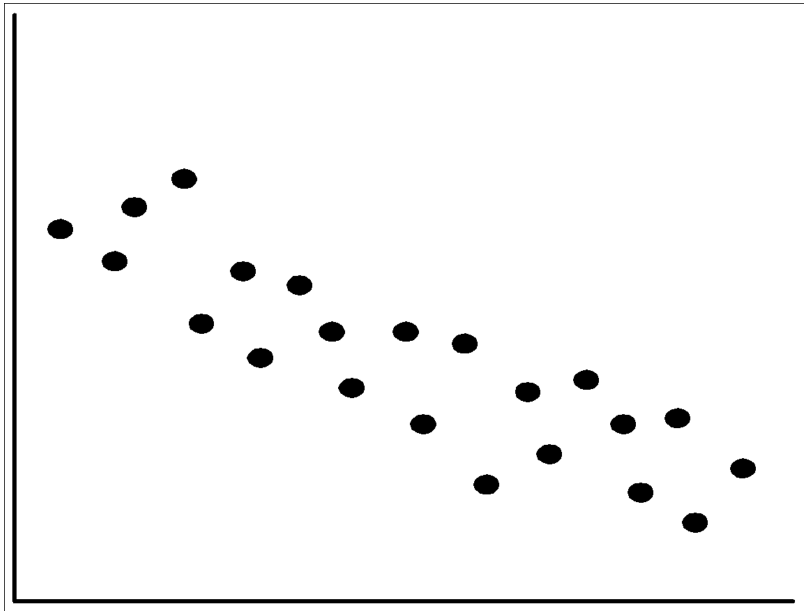
Scatter plot

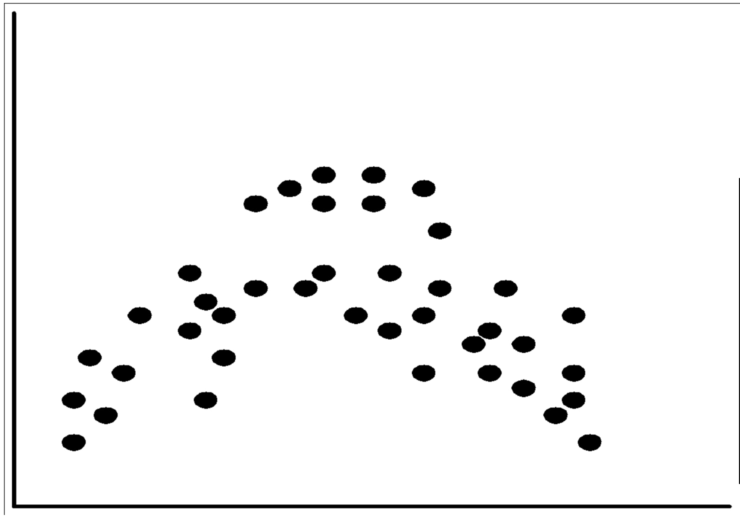
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data

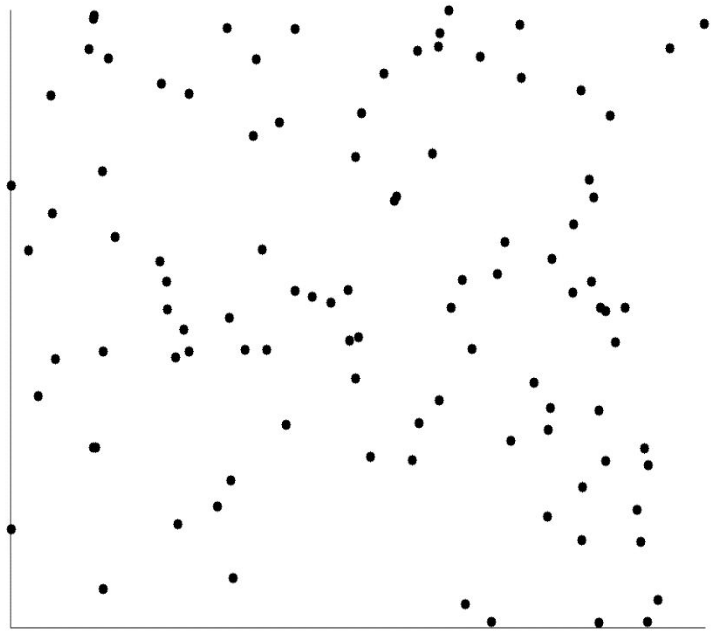


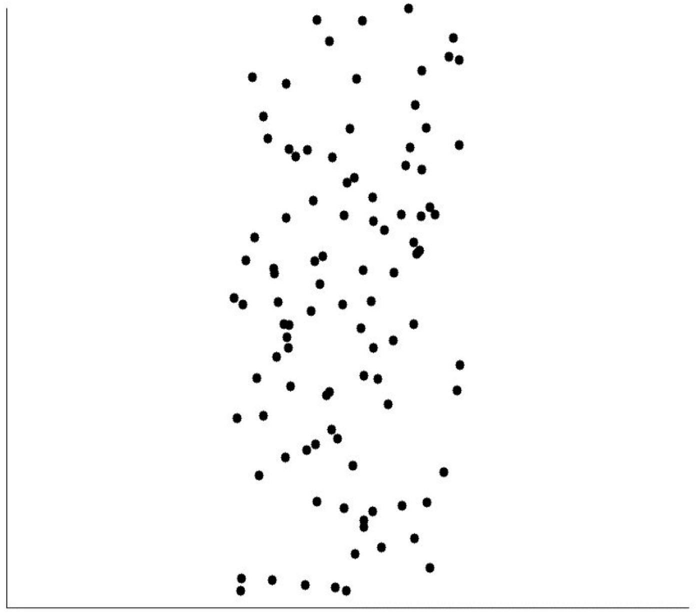




- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data





Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Data visualization: map data onto graphical primitives
  - Measure data similarity
- Above steps are the beginning of data preprocessing.
- Many methods have been developed but still an active area of research.

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu , et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009