Data Mining:

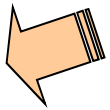
Concepts and Techniques

— Chapter 2 —

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

Getting to Know Your Data

- Data Objects and feature Types



- Basic Statistical Descriptions of Data

Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix
  - 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	play	ball	score	game	win	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	Butter, Bread
3	Butter, Coke, Cookies, Milk
4	Butter, Bread, Cookies, Milk
5	Coke, Cookies, Milk

Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Distribution
  - Centrality and dispersion

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*.
- Data objects are described by **attributes/features**.
- Database rows -> data objects; columns ->feature.

Features

- **Feature**: a data field, representing a characteristic of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Binary
  - Nominal
  - Numeric: quantitative
    - Interval-scaled
    - …

Feature Types

- **Nominal:** categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal feature 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)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings

Discrete vs. Continuous Features

- **Discrete Featrue**
  - 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 features are a special case of discrete features
- **Continuous Feature**
  - Has real numbers as feature values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous features are typically represented as floating-point variables

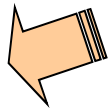
Example: Family Car Data

Example #	Price	Engine Power	Family Car
1	7000	310	no
2	8000	180	no
3	14000	200	no
4	15000	280	yes
5	20000	250	yes
6	20000	340	no
7	21000	290	no
8	22000	300	no
9	25000	260	no
10	27000	285	yes
11	29000	340	no
12	30000	210	no
13	39000	260	no
14	40000	245	no
15	41000	285	no

Example: Family Car Data (multiclass)

Example #	Price	Engine Power	Class
1	7000	310	Family car
2	8000	180	Family car
3	14000	200	Family car
4	15000	280	Family car
5	20000	250	Family car
6	20000	340	Sports car
7	21000	290	Sports car
8	22000	300	Sports car
9	25000	260	Luxury sedan
10	27000	285	Family car
11	29000	340	Sports car
12	30000	210	Luxury sedan
13	39000	260	Luxury sedan
14	40000	245	Luxury sedan
15	41000	285	Sports car

Chapter 2: Getting to Know Your Data

- Data Objects and Feature Types
- Basic Statistical Descriptions of Data 

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
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

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:

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

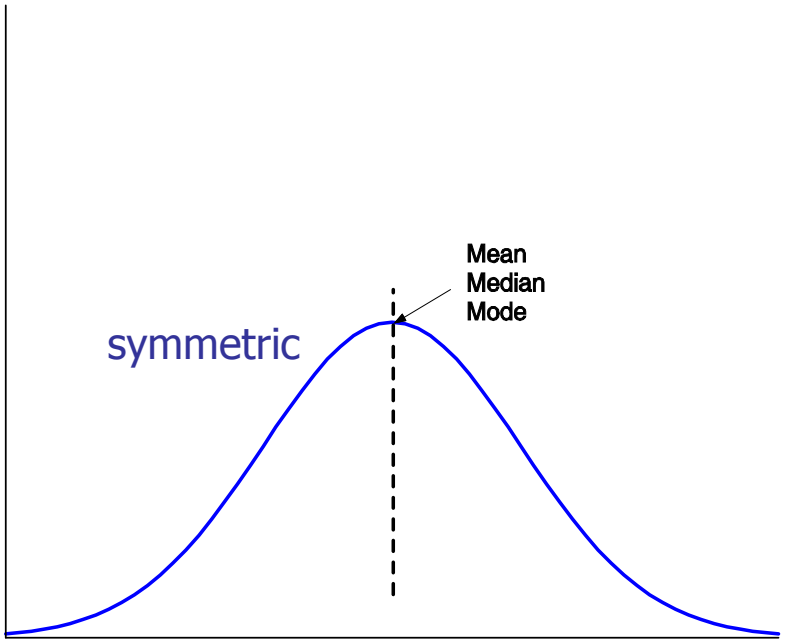
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

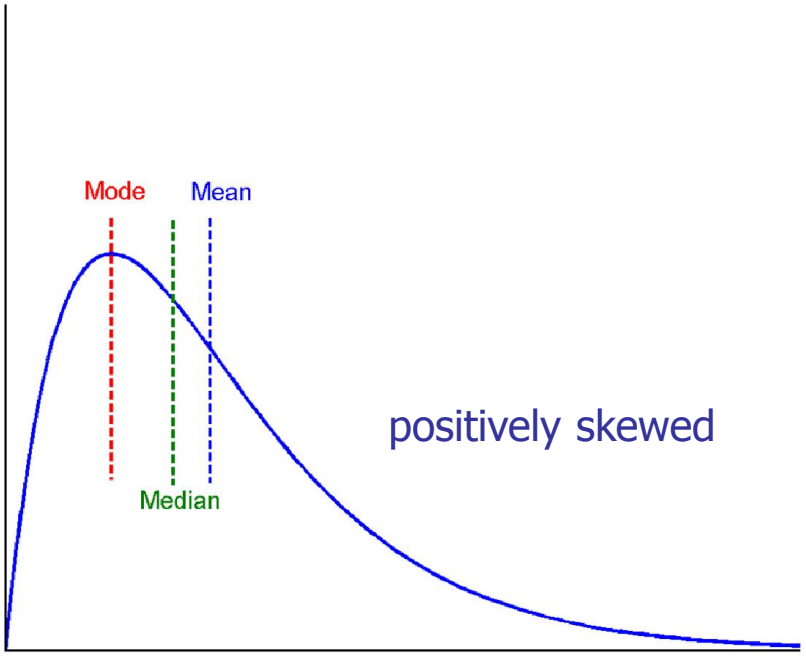
age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

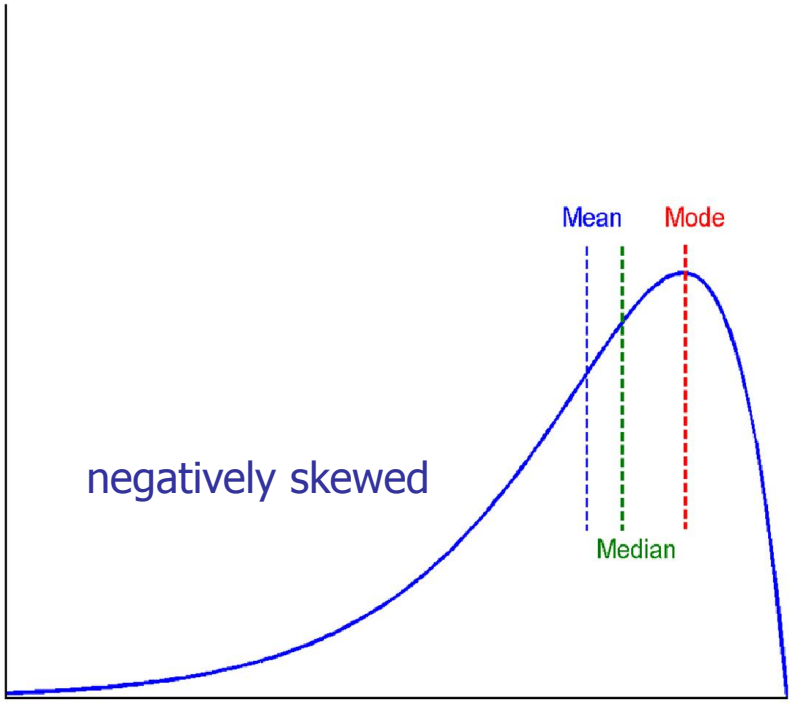
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal

Symmetric vs. Skewed Data

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

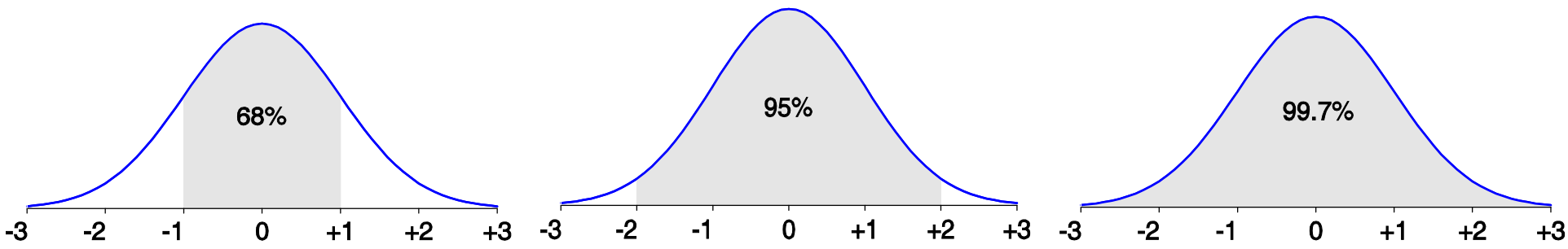






Properties of Normal Distribution Curve

- Galton Board
- 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

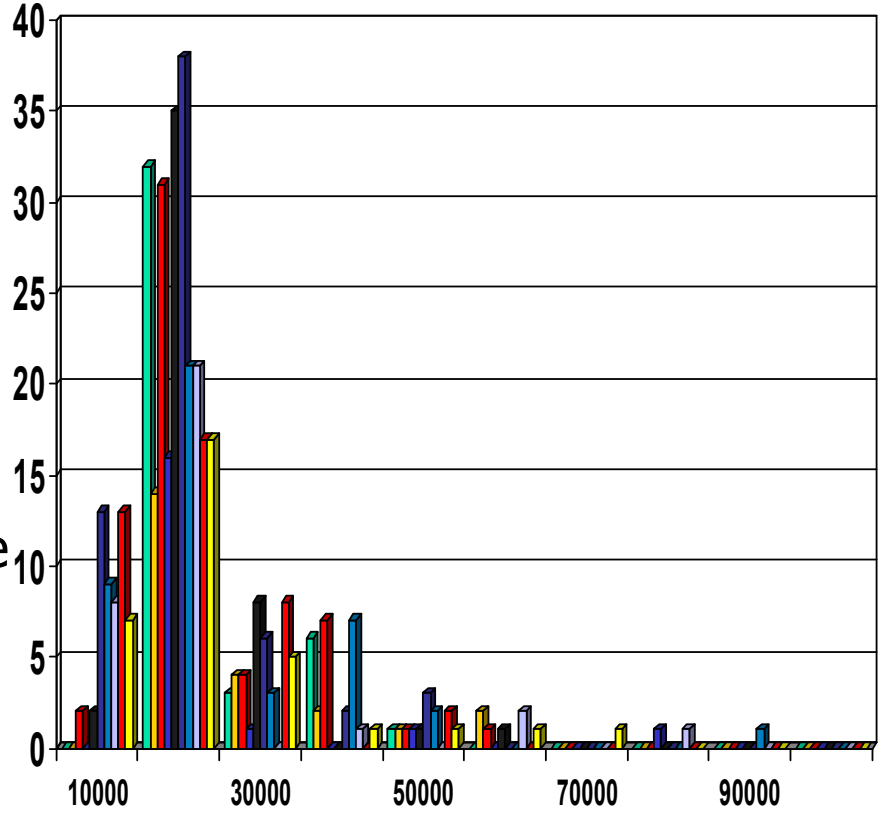


Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **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

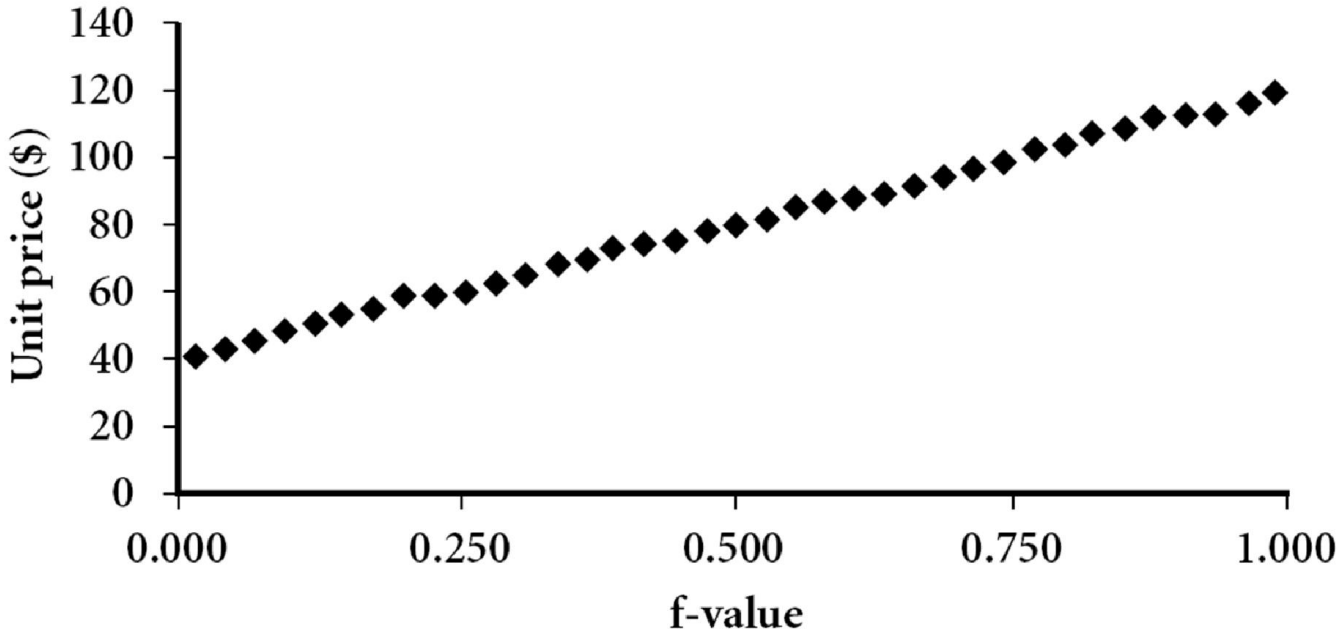
Histogram Analysis

- 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



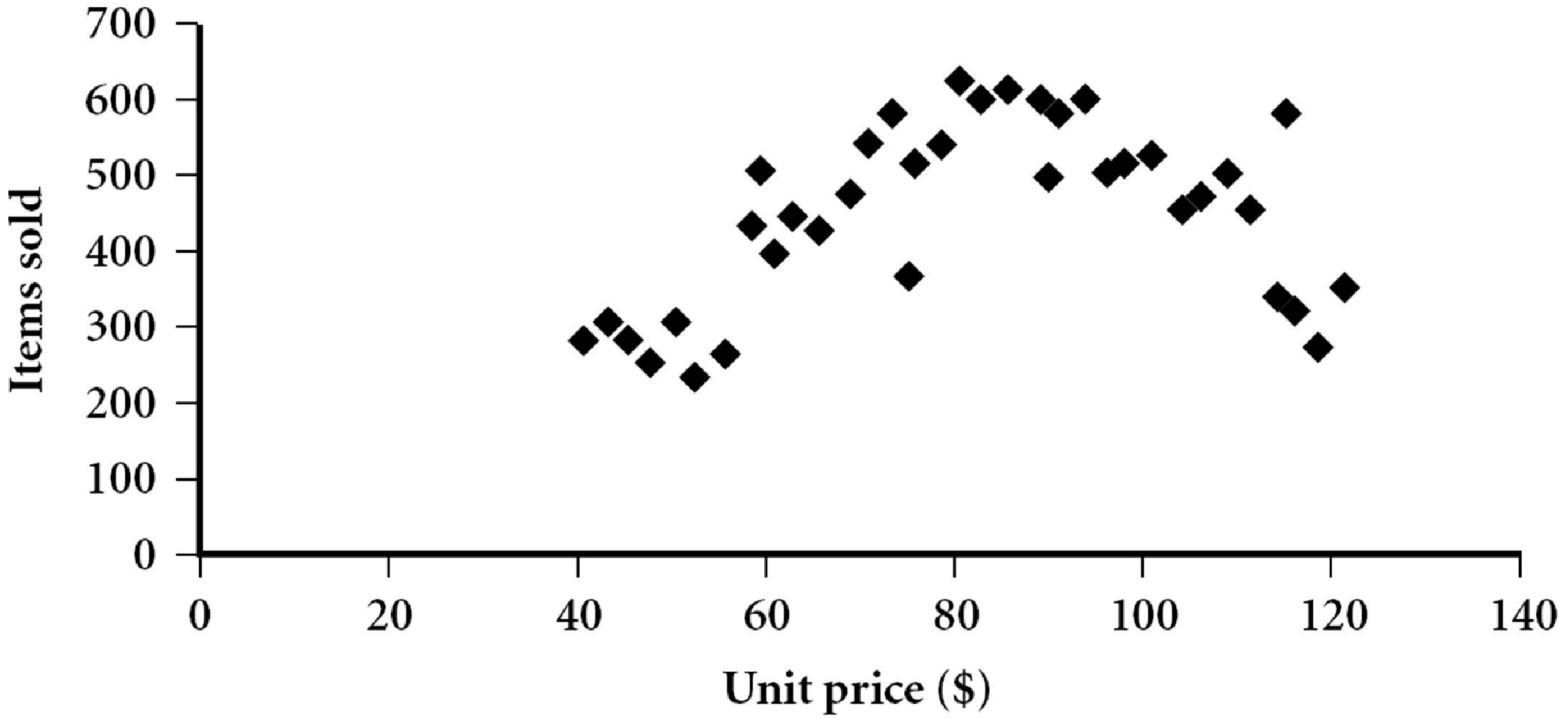
Quantile Plot

- 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

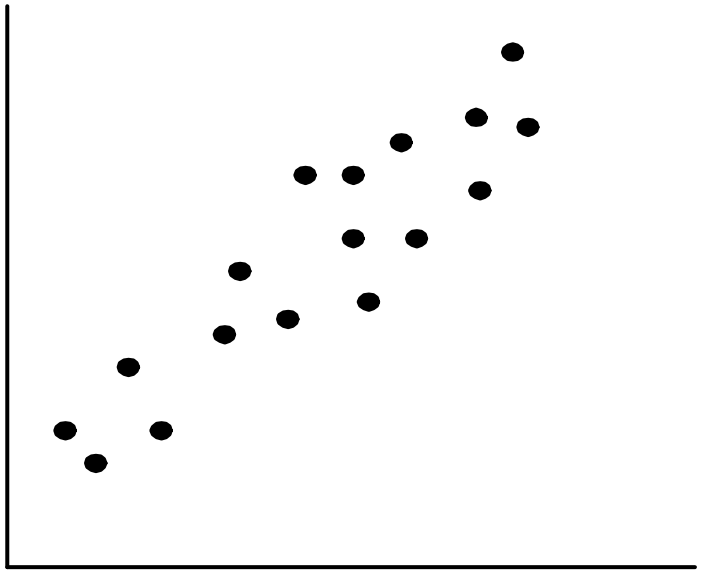


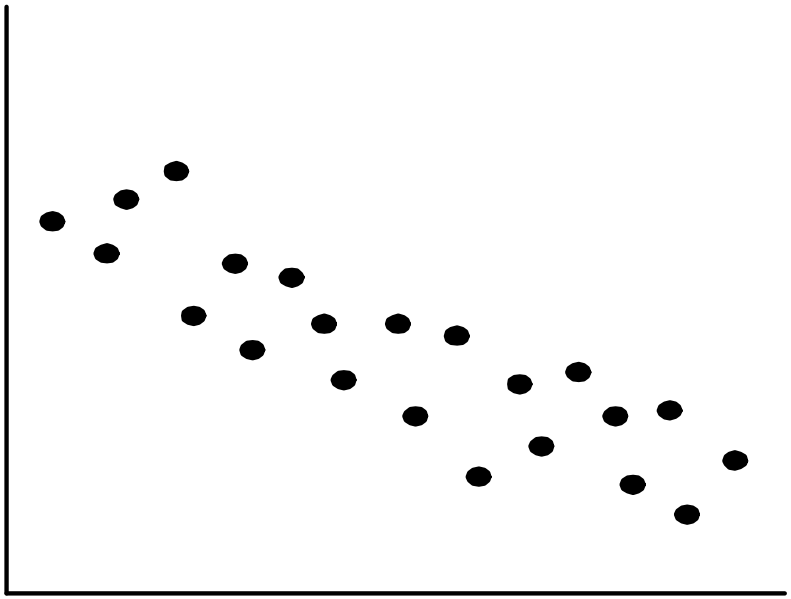
Scatter plot

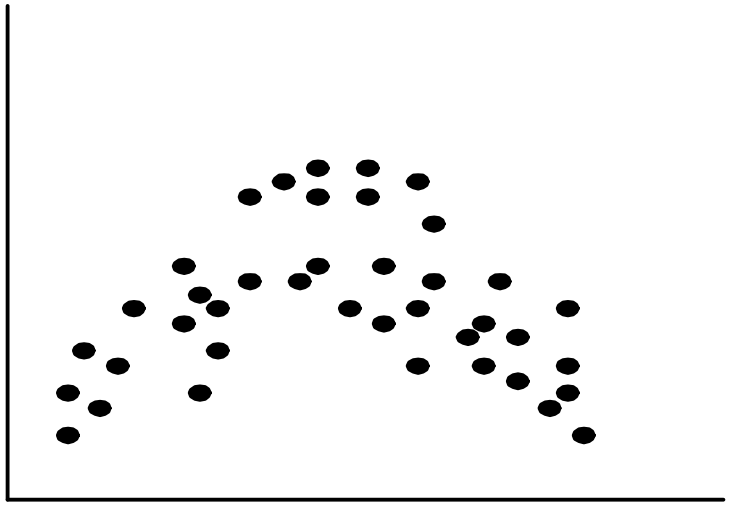
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data







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

Uncorrelated Data