Data Preprocessing

CSC 535/635

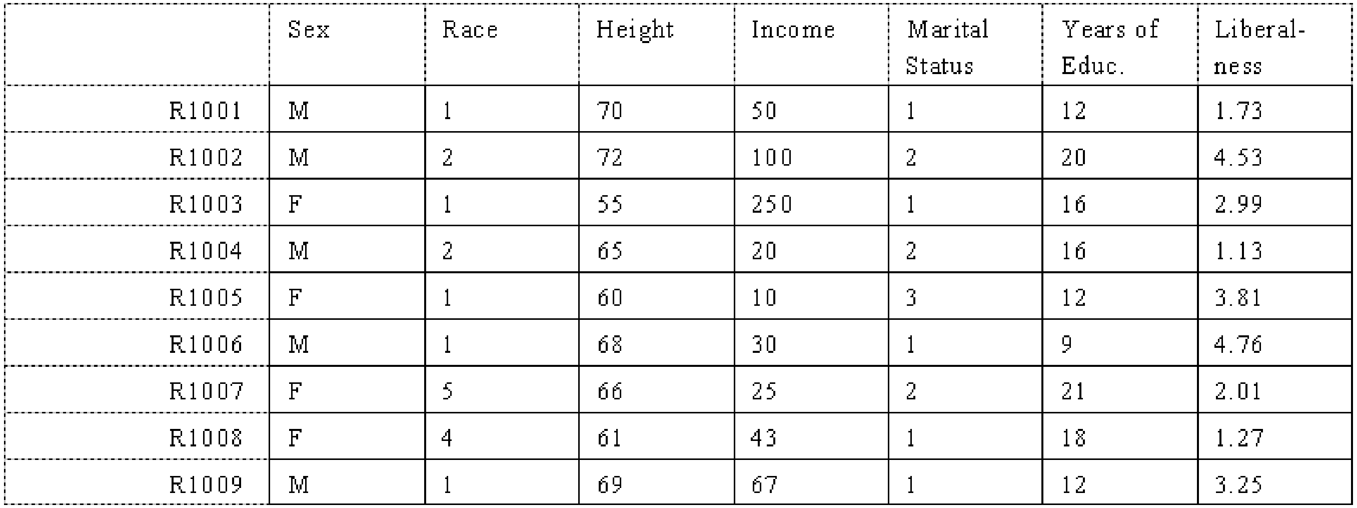
Data Representation

Features (aka attributes)

. . . .

Samples

.
.
.



	Sex	Race	Height	Income	Marital Status	Years of Educ.	Liberal-ness
R1001	M	1	70	50	1	12	1.73
R1002	M	2	72	100	2	20	4.53
R1003	F	1	55	250	1	16	2.99
R1004	M	2	65	20	2	16	1.13
R1005	F	1	60	10	3	12	3.81
R1006	M	1	68	30	1	9	4.76
R1007	F	5	66	25	2	21	2.01
R1008	F	4	61	43	1	18	1.27
R1009	M	1	69	67	1	12	3.25

Types of Variables/Attributes/Data

- Two most common types are:
- Numeric
  - real-value variables or integers variables
  - Ex: age, speed, or length
  - values of a numeric attribute have two important properties:
    - an order relation ($2 < 5$ and $5 < 7$)
    - a distance relation ($d\,[2.3,4.2] = 1.9$)
- Categorical
  - have neither order nor distance relation
  - Ex: eye color, gender, or country of citizenship
  - only support an equality relation (Blue = Blue, or Brown ≠ Black)

Types of Variables - Another classification

- Continuous variables
  - aka quantitative or metric variables
  - values are measured using either an interval scale or a ratio scale
- Discrete variables
  - aka qualitative variables
  - values are measured using one of two kinds of nonmetric scales — nominal or ordinal

Interval Scaled vs. Ratio Scaled Variables

Interval Scale

- The ratio relation does not hold true
- The 0 point is placed arbitrarily, and thus it does not indicate the absence of value
- Ex: temperature F/C
  - 0 F does not mean absence of temperature.
  - 80 F ≠ twice 40 F

Ratio Scale

- The ratio relation holds true
- Ratio scale has an absolute 0 point
- Ex: height, length, salary
  - 6 feet = twice 3 feet

Nominal Scaled vs. Ordinal Scaled Variables

Nominal Scale

- Order-less scale
- May use different symbols, characters, or numbers to represent the different states/values of the variable
- Ex: color, gender, id, customer_type

Ordinal Scale

- Values have an order relation **but not a distance relation**
- Ex: student_class, military_rank, grade

Encoding Numerical Data as Ordinal

- An ordinal variable can be used to encode a numeric variable with a smaller set of values
- Ex: age (with values young, middle aged, and old)
- Ex: income (with values low, middle-class, upper-middle-class, and rich)
- Ex: height (with values short, medium, tall)

Binary Variables

- Has one of two states: 0, 1
- Examples: smoker, owns-house
- Can be considered a special case of nominal variables

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“ ”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., discrepancy between duplicate records

Why is Data Preprocessing Important?

- No quality data, no quality mining results!
- Minimize GIGO (Garbage In Garbage Out)
- Data preparation, cleaning, and transformation comprises the majority of the work in a data mining application

Measures for Data Quality

- Accuracy
- Completeness: not recorded, unavailable
- Consistency: some modified but some not
- Timeliness: timely updates
- Believability
- Interpretability

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization

Data Cleaning

- Importance
  - “Data cleaning is the number one problem in data warehousing”
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Data

Name	Age	Sex	Income	Class
Mike	40	Male	150k	Big spender
Jenny	20	Female	?	Regular
…				

- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
- Missing data may need to be inferred

How to Handle Missing Data?

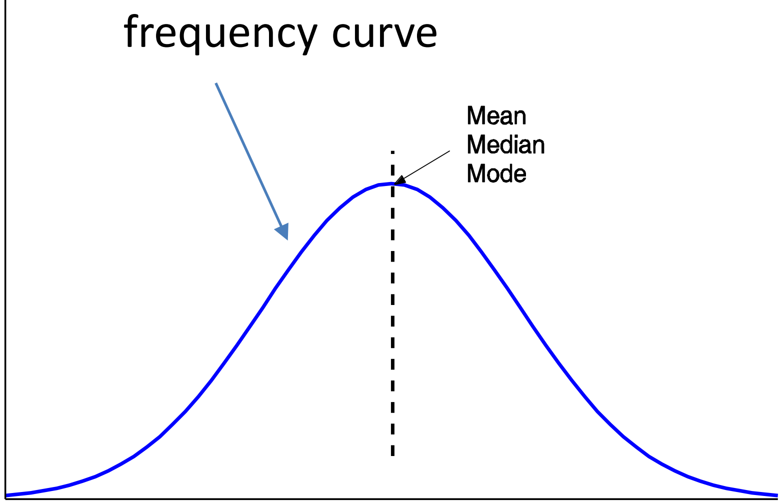
- Ignore the tuple
- Fill in missing values manually: tedious & infeasible
- Fill in it automatically with
  - a global constant : e.g., “unknown”
    - may confuse DM algorithm
  - the attribute mean or median
  - the attribute mean for all samples belonging to the same class
  - the most probable value: using inference-based tools such as Bayesian formula or decision tree induction

Measuring the Central Tendency

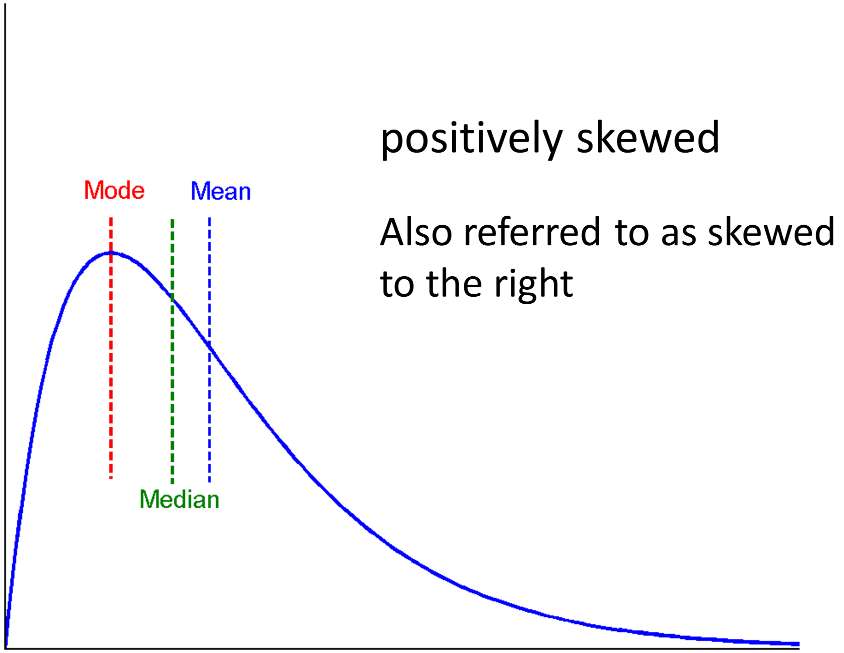
- Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Weighted arithmetic mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - Trimmed mean: chopping extreme values
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal, multimodal
  - Empirical formula: $mean - mode \approx 3 \times (mean - median)$

Symmetric vs. Skewed Data

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

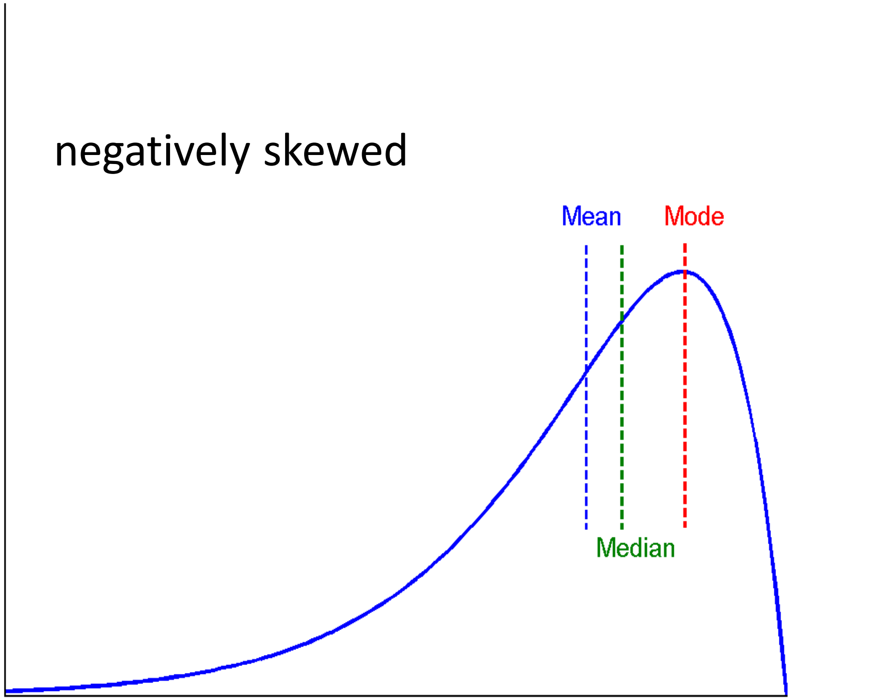


frequency curve



positively skewed

Also referred to as skewed to the right



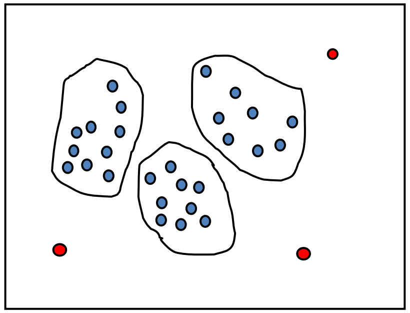
negatively skewed

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - etc
- Other data problems which requires data cleaning
  - duplicate records, incomplete data, inconsistent data

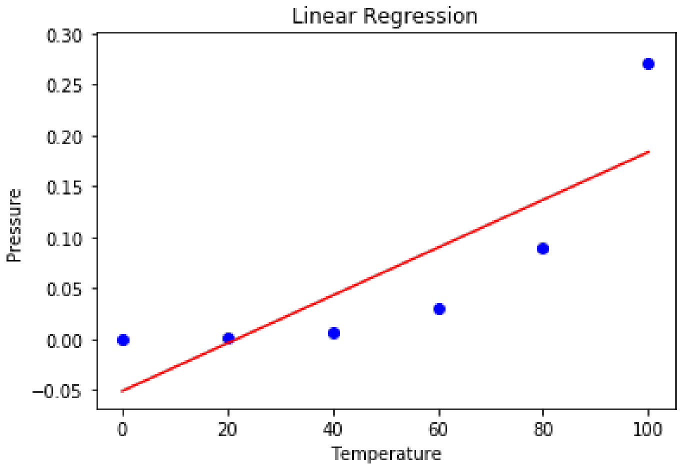
How to Handle Noisy Data?

- Binning method
  - first sort data and partition into (equal-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
  - detect and remove outliers



How to Handle Noisy Data? (cont.)

- Regression
  - smooth by fitting the data into regression functions



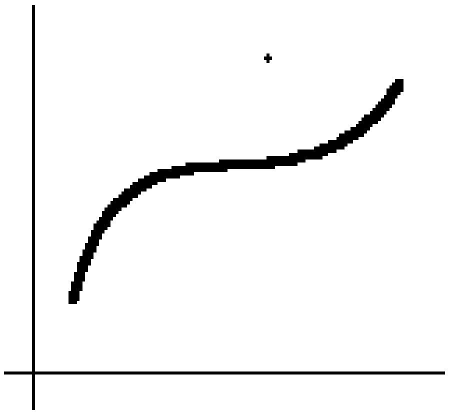
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into (equal-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Outlier Removal

- Data points inconsistent with the majority of data
- Different outliers
  - Valid: CEO's salary,
  - Noisy: One's age = 200, widely deviated points
- Removal methods
  - Clustering
  - Curve-fitting



Data Integration

- Data integration:
  - combines data from multiple sources
- Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources can be different, e.g., different scales, metric vs. British units
- Removing duplicates and redundant data

Data Transformation

- Transforming the data into a form appropriate for mining.
- Methods:
  - Smoothing: remove noise from data
  - Aggregation: summarization
  - Normalization: scaled to fall within a small, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Attribute/feature construction
    - New attributes constructed from the given ones

Data Transformation: Normalization

- Given numeric attribute A with values v_1, v_2, …, v_N
- **min-max normalization:** to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- **z-score normalization**

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- **decimal scaling normalization**

$$v' = \frac{v}{10^j}$$

Where j is the number of digits in the data value with the largest absolute value

Min-Max Normalization – Example

- Let income range from $12,000 and $98,000.
- Normalize to the range [0.0, 1.0].
- Using min-max normalization, an income value of $73,600 is transformed to

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0.0) = 0.716$$

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Z-Score Normalization – Example

- Given mean = $54,000 and standard deviation = $16,000
- Using z-score normalization, an income value of $73,600 is transformed to

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

Decimal Scaling Normalization – Example

- Suppose values of A range from -4986 to 7845
- To normalize by decimal scaling, divide by ______
- -4986 normalizes to -0.4986
- 7845 normalizes to 0.7845

$$v' = \frac{v}{10^j}$$

Where *j* is the number of digits in the data value with the largest absolute value

Variation of Z-Score Normalization

- Z-score normalization $$v' = \frac{v - mean_A}{stand\_dev_A}$$
- stand_dev = $\sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$
- stand_dev is sensitive to outliers
- A variation of z-score normalization that is more robust to outliers replaces $(x_i - \bar{x})^2$ by $|x_i - \bar{x}|$
- Mean absolute deviation $$M.A.D. = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{n}$$
- Z-score using MAD $$v' = \frac{v - mean_A}{M.A.D.}$$

Data Reduction

- Data may be too big to work with
- Data reduction
  - Obtain a reduced representation of the data set that is smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
  - Dimensionality reduction: e.g., remove unimportant attributes
  - Reducing the number of attribute values
  - Reducing the number of tuples

Dimensionality Reduction

- Feature selection (i.e., attribute subset selection)
  - Select a minimum set of attributes (features) that is sufficient for the data mining task
- Heuristic methods (due to exponential # of choices)
  - step-wise forward selection
  - step-wise backward elimination
  - combining forward selection and backward elimination

Dimensionality Reduction

- Other techniques include
- Principle component analysis
- Wavelet transforms

Reducing the Number of Attribute Values

- Some algorithms such as decision trees compare different attribute’s values
- Approaches
  - Binning: replace with bin means, medians, boundaries
  - Discretization: discretize a numeric attribute (income) into a small number of intervals, then map each interval to a discrete symbol (low, middle-income, high)

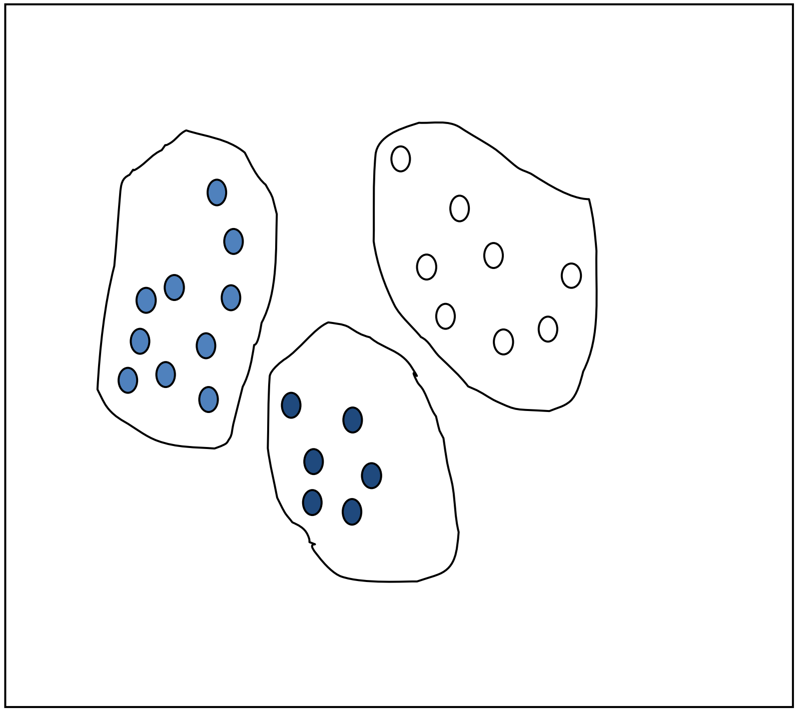
Reducing The Number of Tuples – Sampling

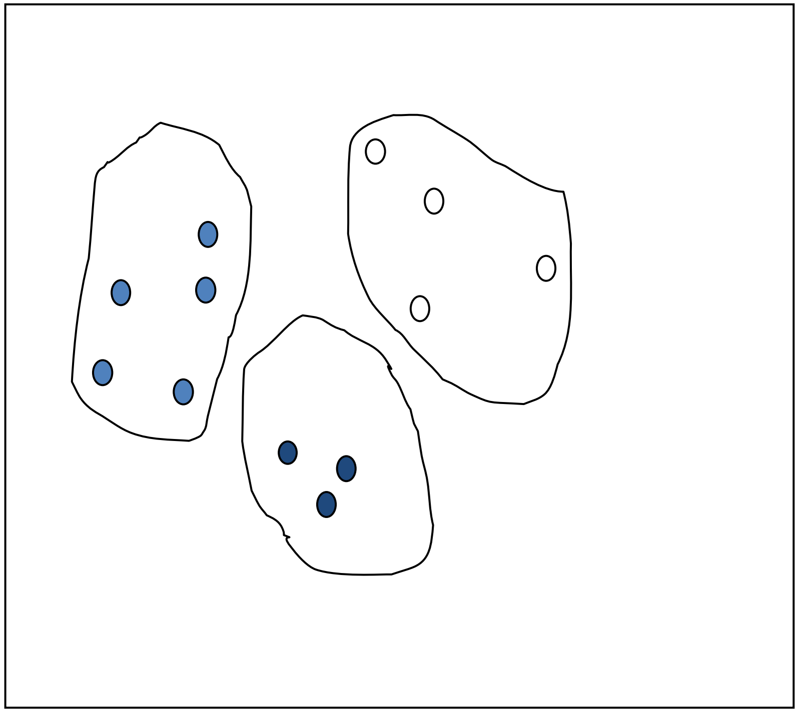
- Sampling: obtaining a small sample s to represent the whole dataset N
- Simple random sampling may have poor performance if the dataset is skewed
- Choose a representative subset of the data
  - Stratified sampling: approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data

Stratified Sampling

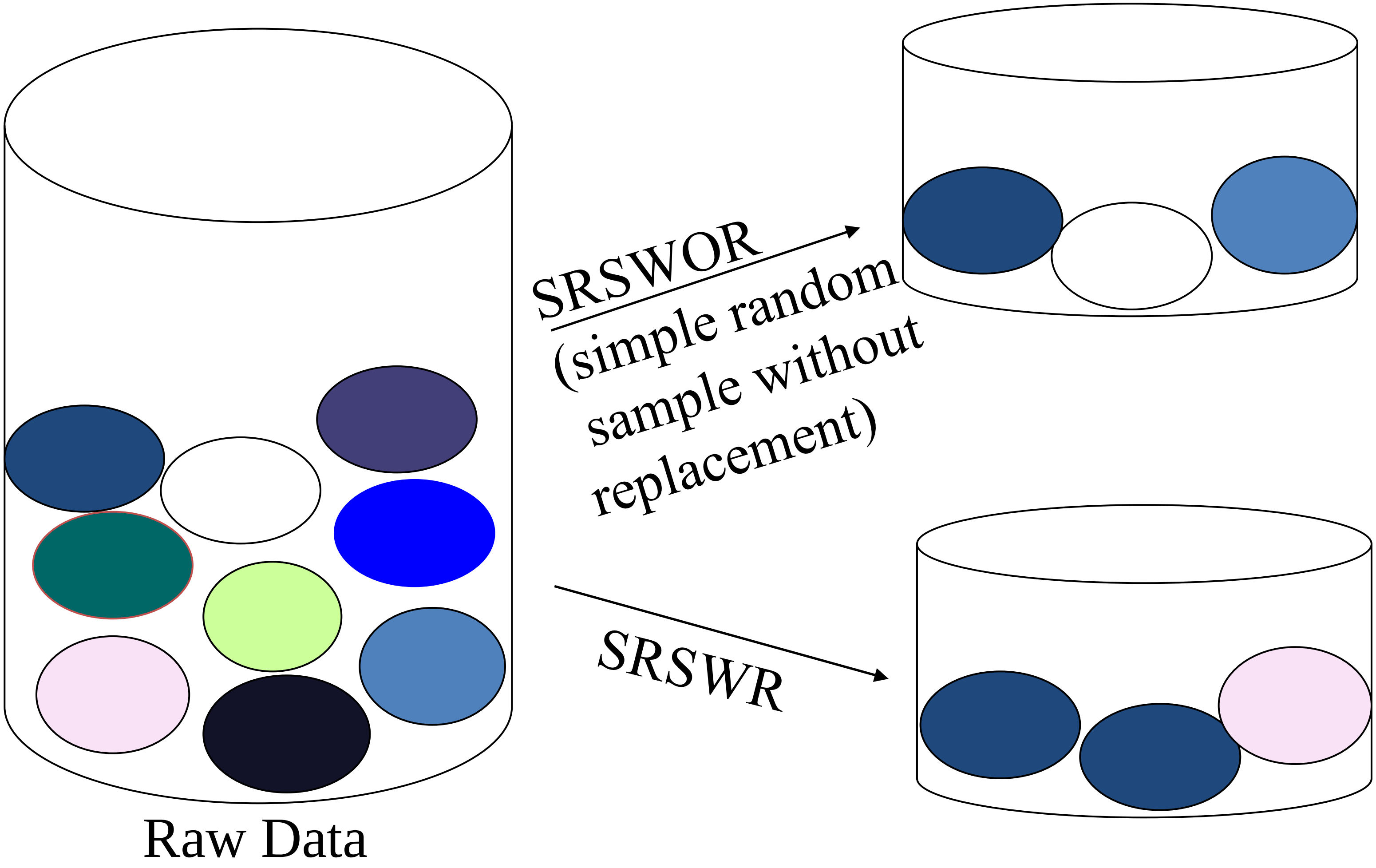
Raw Data

Stratified Sample





Sampling: with or without Replacement

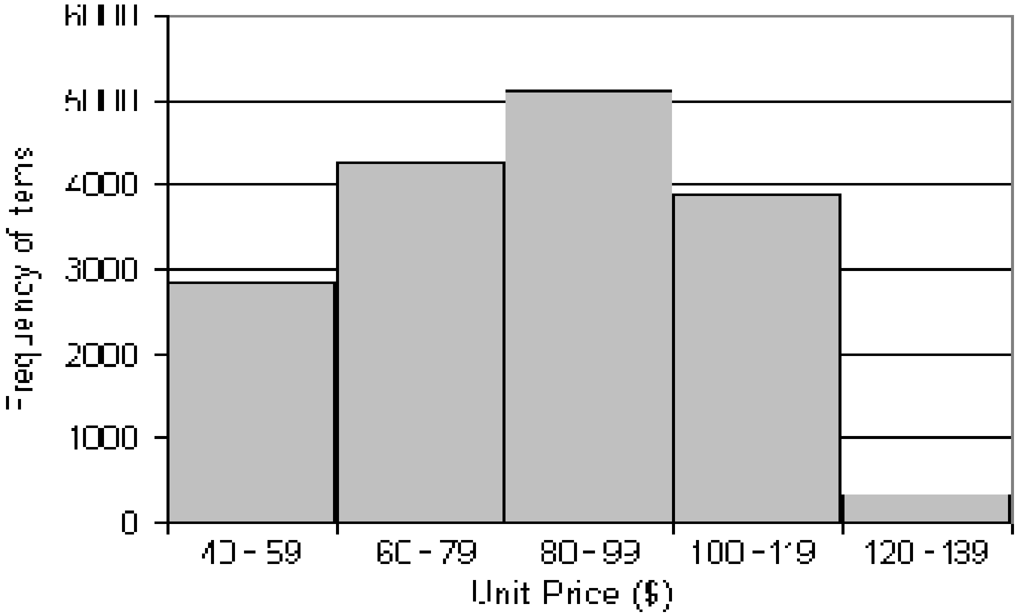


Discretization

- Three types of attributes:
  - Nominal — values from an unordered set
  - Ordinal — values from an ordered set
  - Continuous — real numbers
- Discretization:
  - discretize the range of a continuous attribute
- Some techniques:
  - Binning methods
  - Clustering-based methods
  - Entropy-based methods (Decision Trees)
  - Histogram analysis

Histogram Analysis

- Graph displays of basic statistical class descriptions
  - Frequency histograms
  - A graphical method
  - Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data

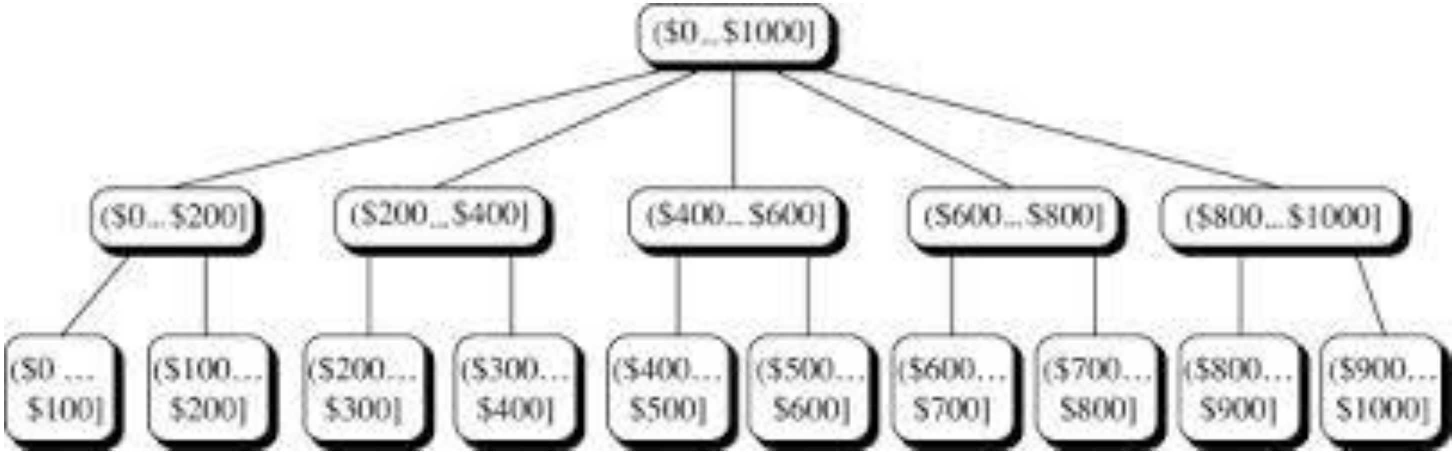


Discretization and Concept Hierarchies

- Discretization
  - Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals.
  - Interval labels can then be used to replace actual data values
- Concept hierarchies
  - Reduce the data by collecting and replacing low level concepts (numeric values for age) by higher level concepts (such as young, middle-aged, or senior)
  - Things can be done at different levels
  - Concept hierarchies for **nominal attributes**, where attributes such as street can be generalized to higher-level concepts, like city, state, or country

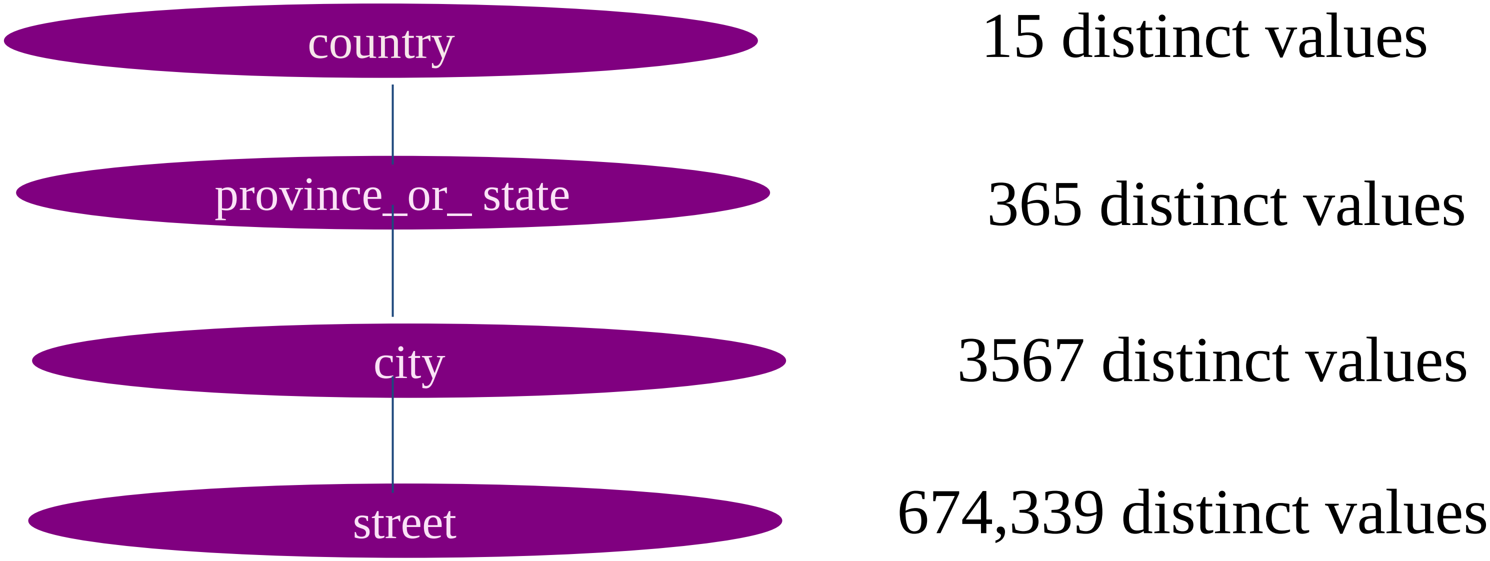
Concept Hierarchy

- Concept hierarchy for attribute price



Concept Hierarchy

- Concept hierarchy for attribute street



country

15 distinct values

province_or_ state

365 distinct values

city

3567 distinct values

street

674,339 distinct values

Summary

- Data preparation is a big issue for data mining
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- Many methods have been proposed but still an active area of research

Measuring the Dispersion of Data

- Variance, standard deviation, range, quartiles, interquartile range, the five-number summary, and boxplots
- They can be used to find outliers
- Variance and std. dev. indicate **how spread out a data distribution is**
  - low *σ* means the data tend to be close to the mean
  - large *σ* indicates that the data are spread out
- Variance and standard deviation (*sample: s, population: σ)*
  - Variance

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

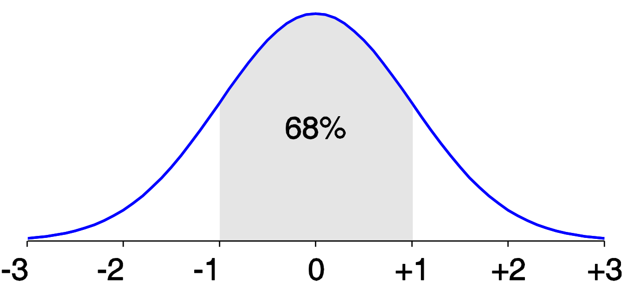
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} {x_i}^2 - \mu^2$$

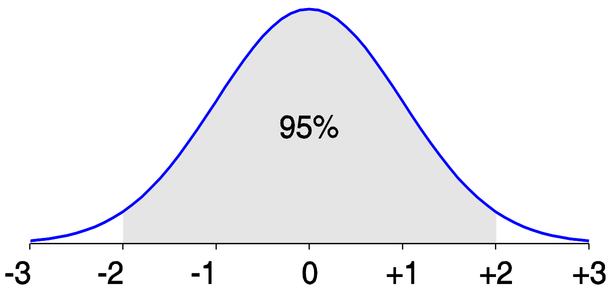
  - Standard deviation *s (or σ)* is the square root of variance s^2 $($*or* $\sigma^2)$
- Remember: Mean (sample: $\bar{x}$, population: μ)

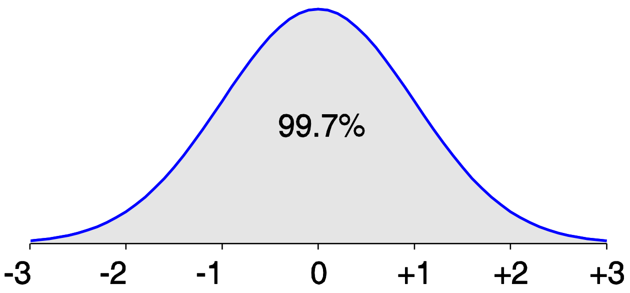
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

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

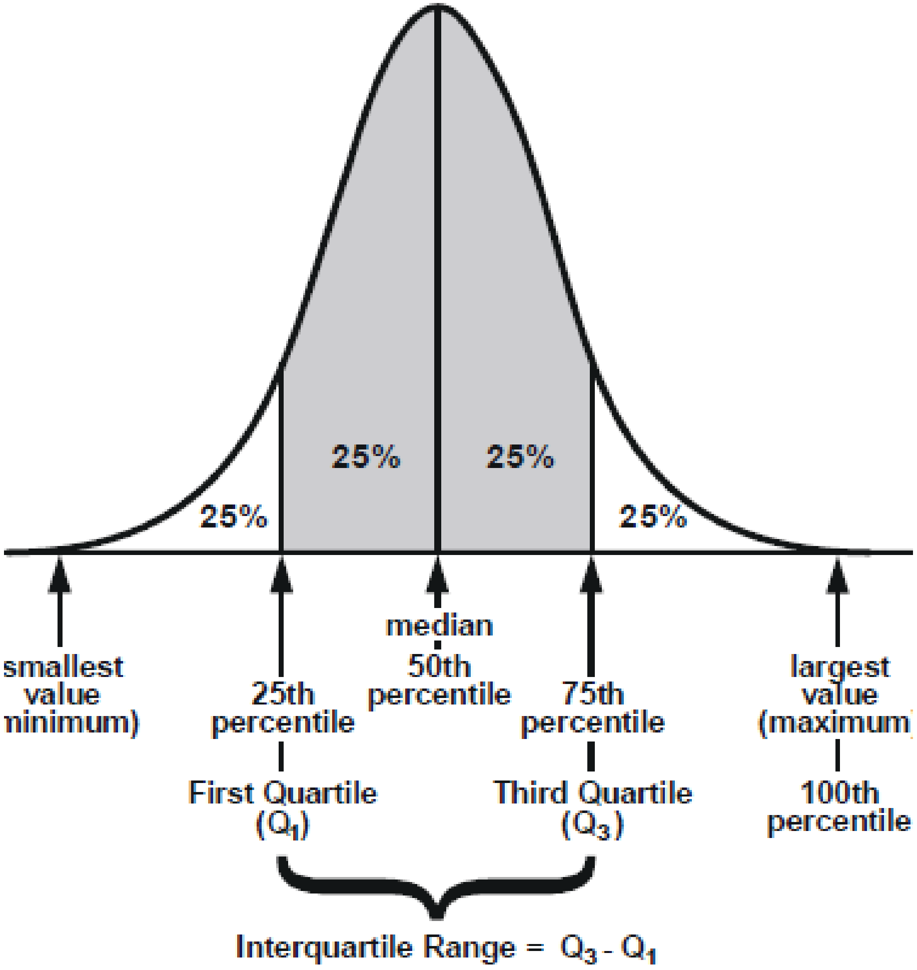






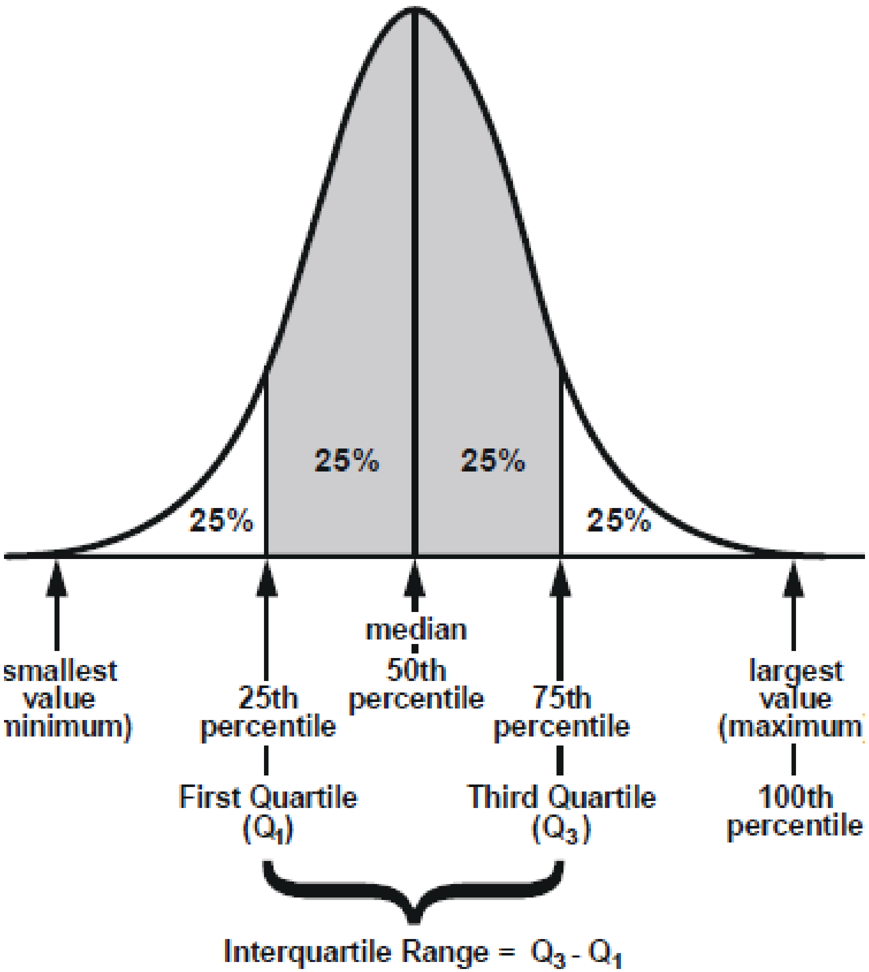
Measuring the Dispersion of Data (cont.)

- Let $x_1, x_2, \ldots, x_N$ be the values of a numeric attribute, X.
- range(X) = max – min
- Assume that the values of X are sorted in increasing order
- The **kth q-quantile** is the value x s.t. at most k/q of the data values are less than x and at most (q-k)/q of the data values are more than x
- **Quantiles** are values in X that allow us to break the data distribution into equal-size consecutive sets
- When we break X into 4 equal parts, the quantiles are called **quartiles**
- When we break X into 100 equal parts, the quantiles are called **percentiles**



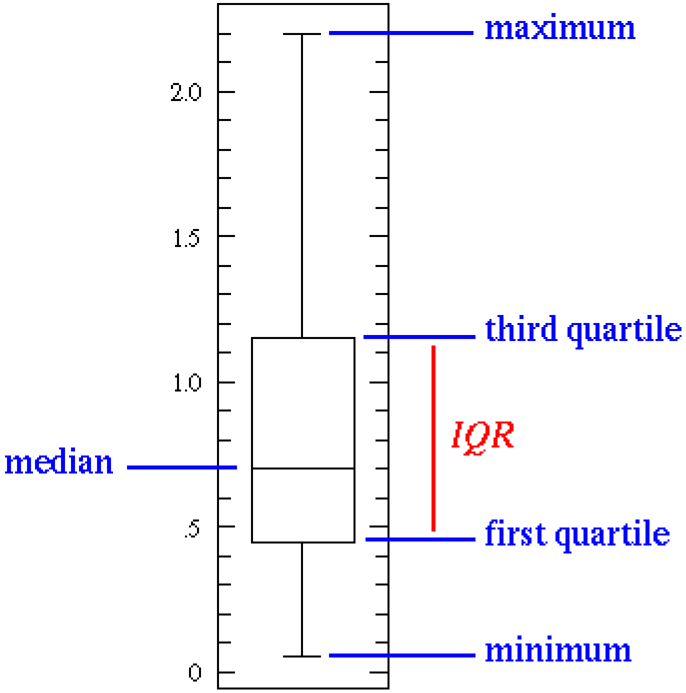
Measuring the Dispersion of Data (cont.)

- Quartiles:
  - Q_1 (25th percentile)
  - Q2 (50th percentile) = median
  - Q_3 (75th percentile)
- Inter-quartile range:
  IQR = $Q_3 – Q_1$
- Five number summary: min, Q_1, M, Q_3, max
- Outlier: usually, a value higher/lower than 1.5 x IQR above/below Q_3/Q_1



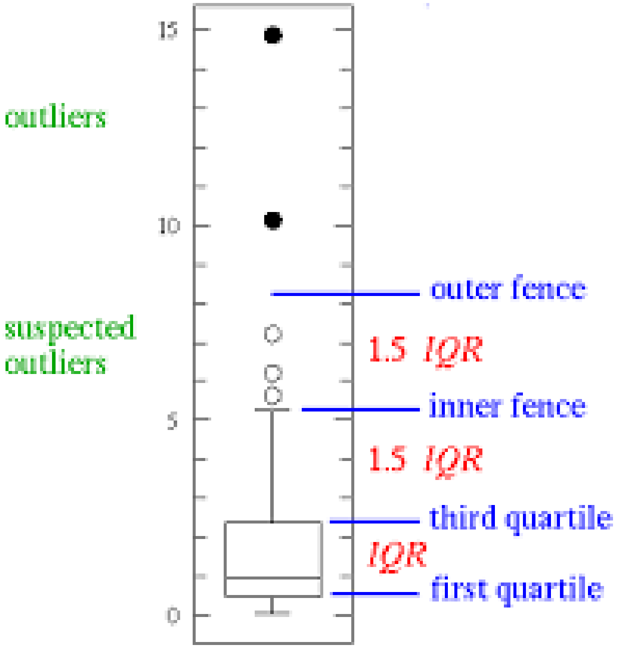
Boxplots

- A way of visualizing a distribution
- Shows five-number summary: min, Q1, M, Q3, max
- Boxplot:
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles -- the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum

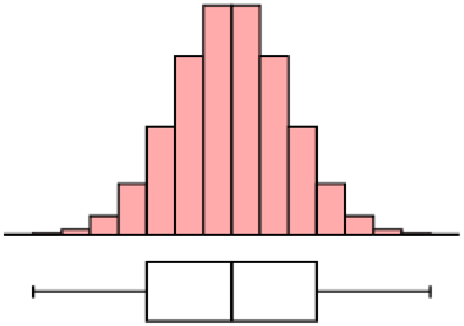


Boxplots (cont.)

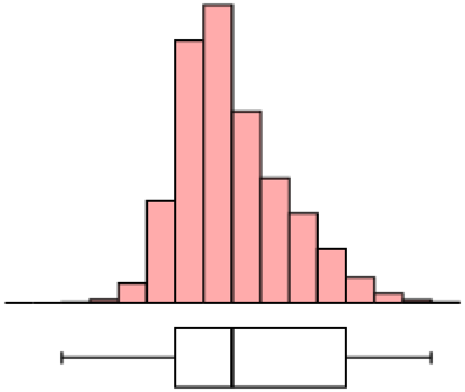
- Outlier: usually, a value higher/lower than 1.5 x IQR
- Whiskers extend to extreme lows only if less/higher than 1.5 x IQR Q_1/Q_3
- Otherwise, terminate at terminate at 1.5 x IQR beyond Q_1/Q_3
- Remaining points are plotted individually



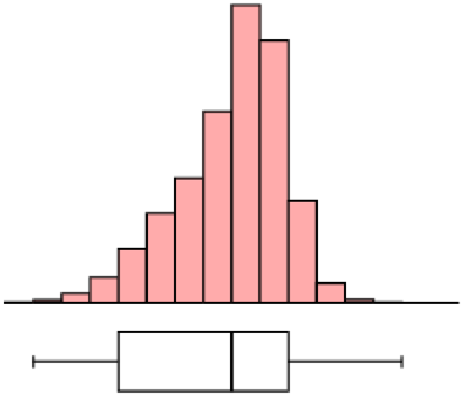
Boxplots of Symmetric & Skewed Data



Symmetric



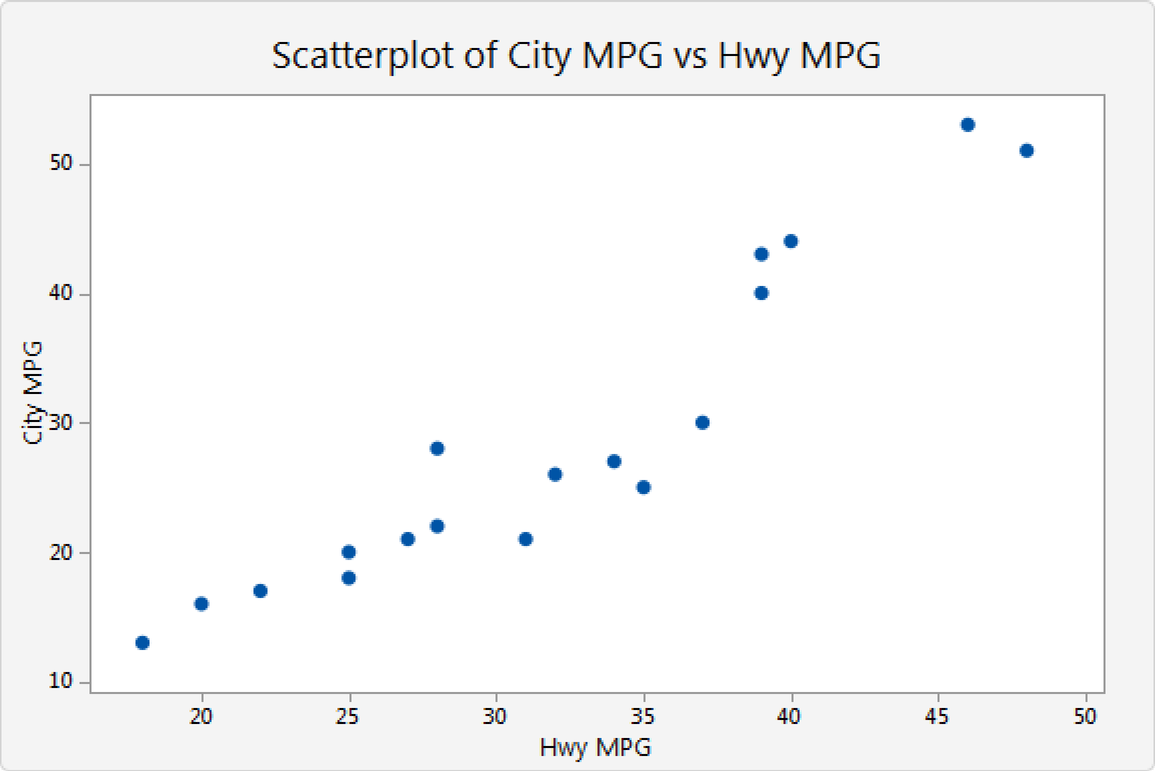
Skewed right
(positive)



Skewed left
(negative)

Scatter plots

- 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


Scatterplot of City MPG vs Hwy MPG
City MPG
Hwy MPG