

ECLT 5810

Data Preprocessing

Prof. Wai Lam

Why Data Preprocessing?

- Data in the real world is imperfect
 - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - noisy: containing errors or outliers
 - inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining results!
 - Quality decisions must be based on quality data
 - Data warehouse needs consistent integration of quality data

Types of Data Sets

- Relational records

NAME	AGE	INCOME	CREDIT RATING
Mike	<= 30	low	fair
Mary	<= 30	low	poor
Bill	31..40	high	excellent
Jim	>40	med	fair
Dave	>40	med	fair
Anne	31..40	high	excellent

- Transaction data

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

- Data matrix, e.g., numerical matrix
- Document data:
 - 3 – text documents: term-frequency vector

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”
 - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
 - marital status, occupation, ID numbers, zip codes
 - **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)
 - **Ordinal**
 - Values have a meaningful order (ranking) but magnitude between successive values is not known.
- 6
- *Size* = {*small, medium, large*}, grades

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
 - Measured on a scale of **equal-sized units**
 - Values have order
 - e.g., *calendar dates*
 - No true zero-point

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
 - 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):

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

- Weighted arithmetic mean:
- Trimmed mean: chopping extreme values

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

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

- Median:

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

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

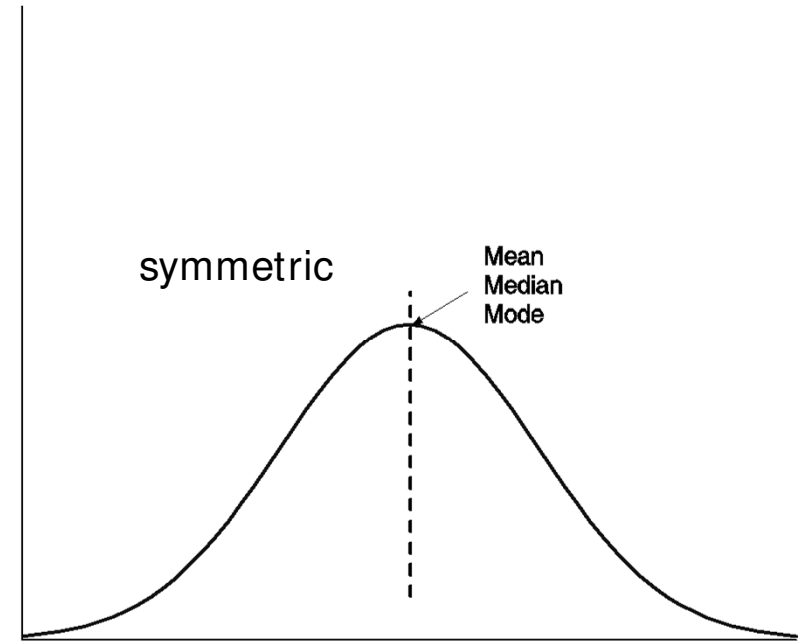
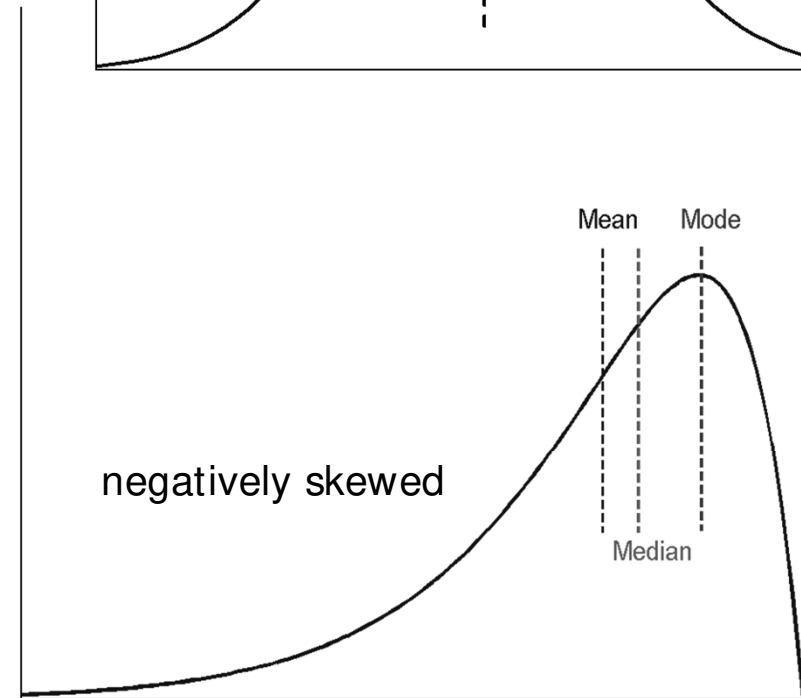
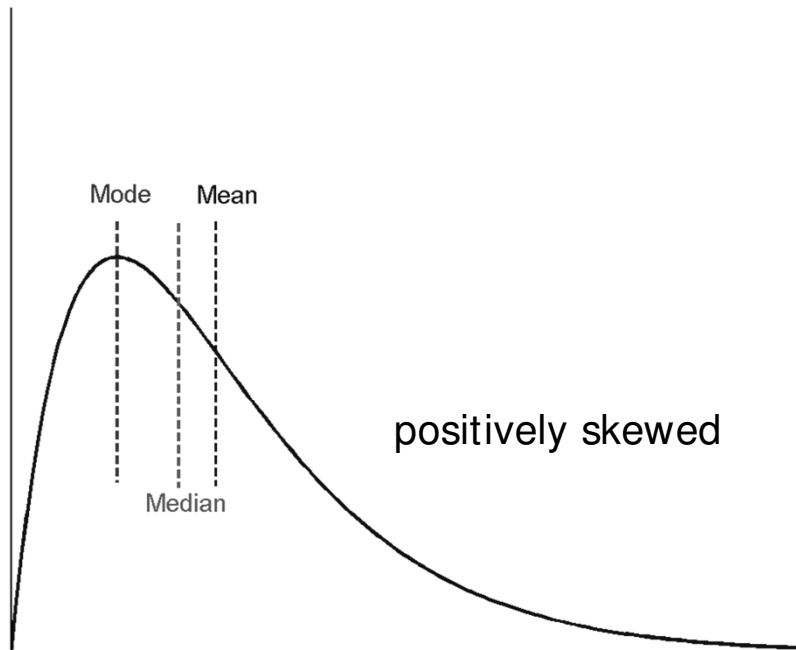
- Mode

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

- Empirical formula: $mean - mode = 3 \times (mean - median)$

Symmetric vs. Skewed Data

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



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

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

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

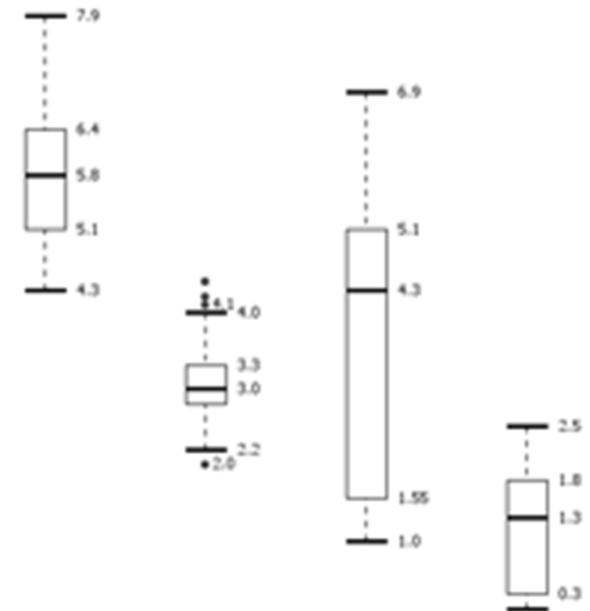
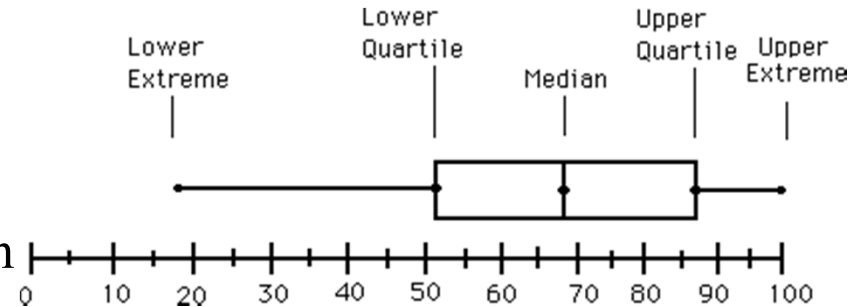
Boxplot Analysis

- **Five-number summary** of a distribution

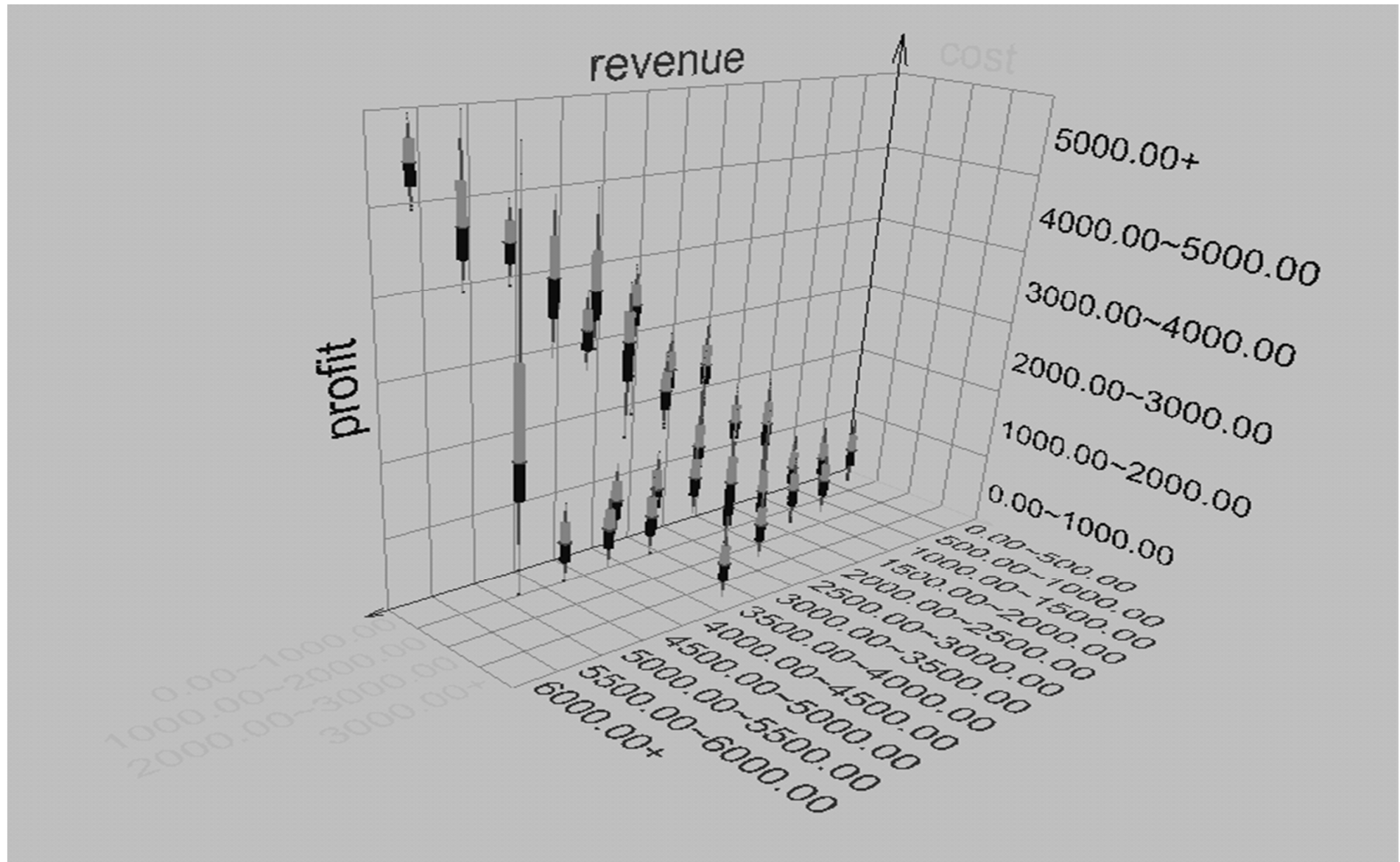
- Minimum, Q1, Median, Q3, Maximum

- **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

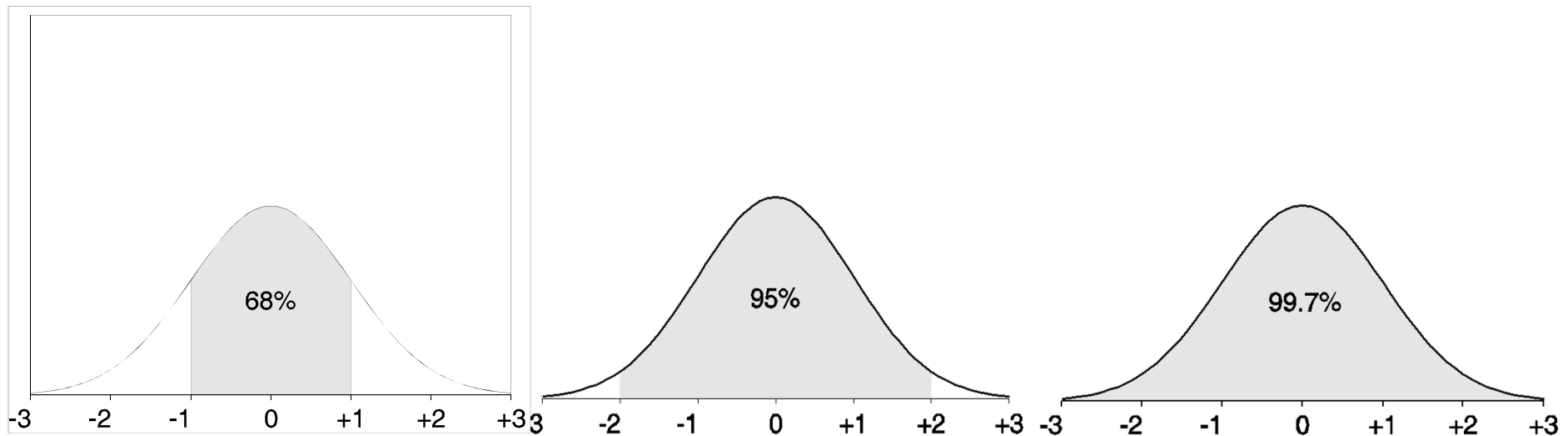


Visualization of Data Dispersion: 3-D Boxplots



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

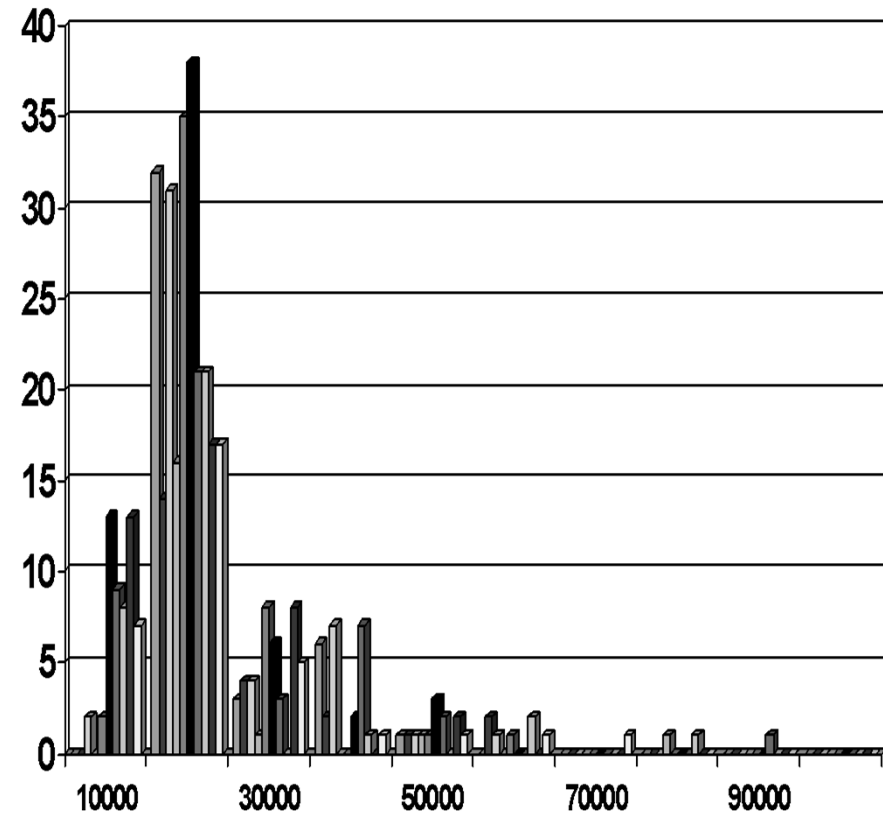


Graphic Displays of Basic Statistical Descriptions

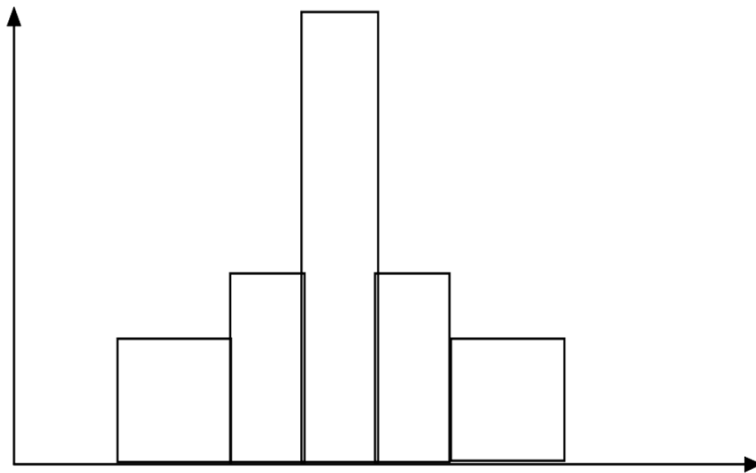
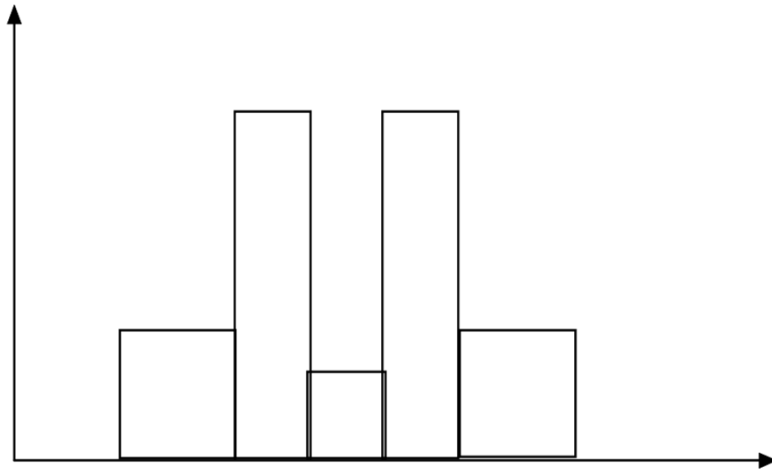
- **Boxplot:** graphic display of five-number summary
- **Histogram:** x-axis are values, y-axis repres. frequencies
- **Quantile plot:** each value x_i is paired with f_i indicating that approximately $100 f_i \%$ of data are $\leq x_i$
- **Scatter plot:** each pair of values is a pair of coordinates 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



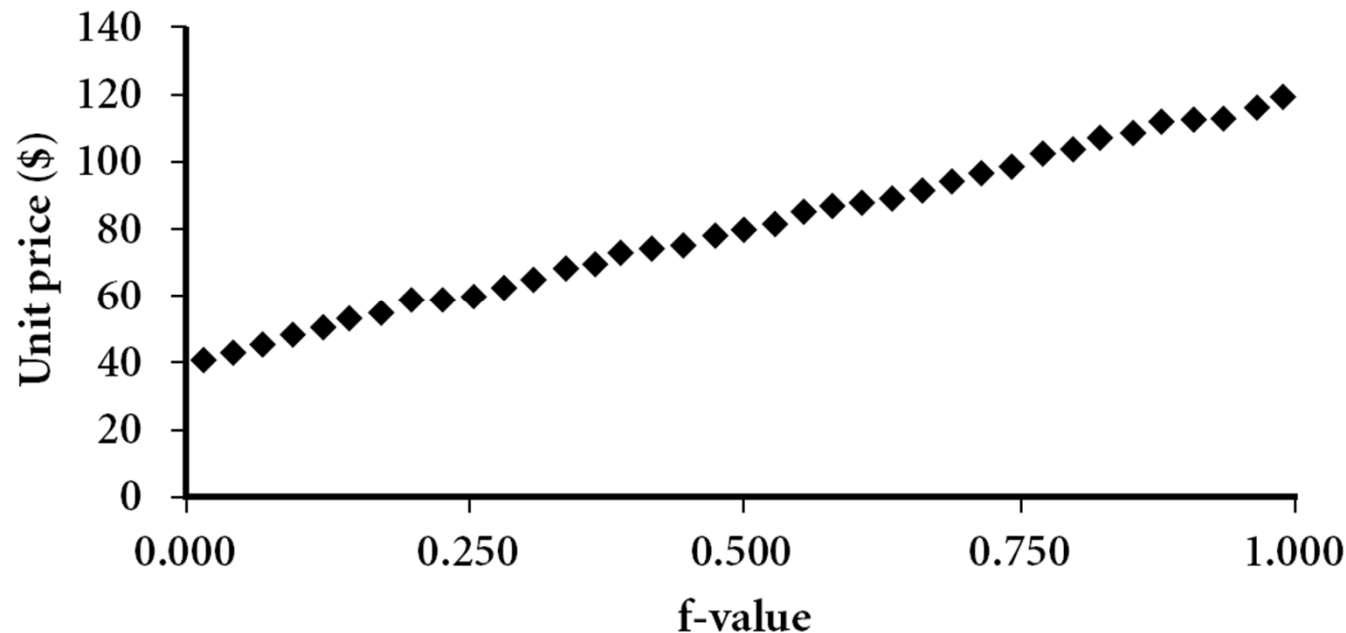
Histograms Often Tell More than Boxplots



- 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

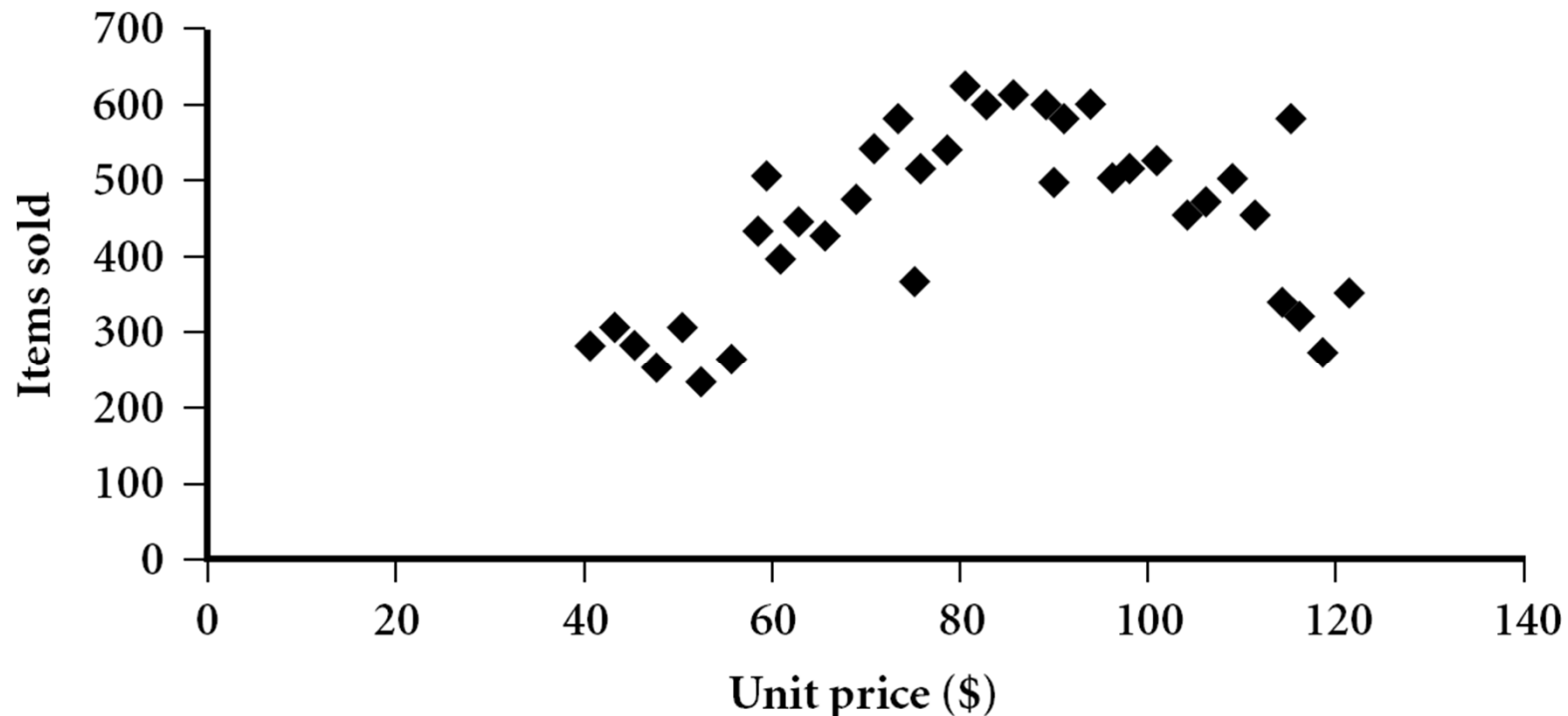
Quantile Plot

- 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

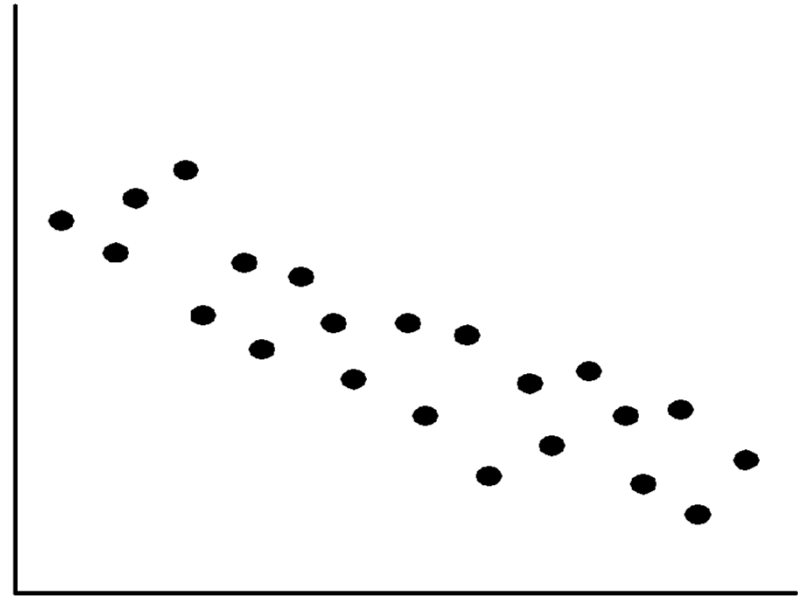


Scatter plot

- 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

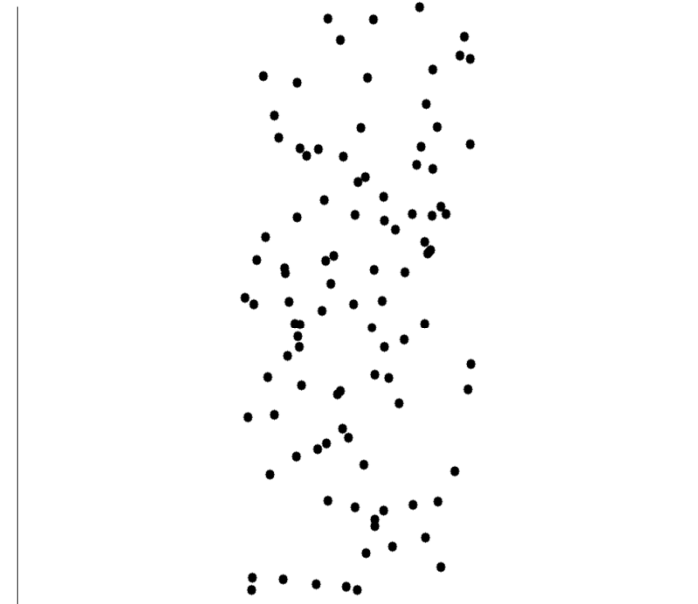


Positively and Negatively Correlated Data



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

Uncorrelated Data

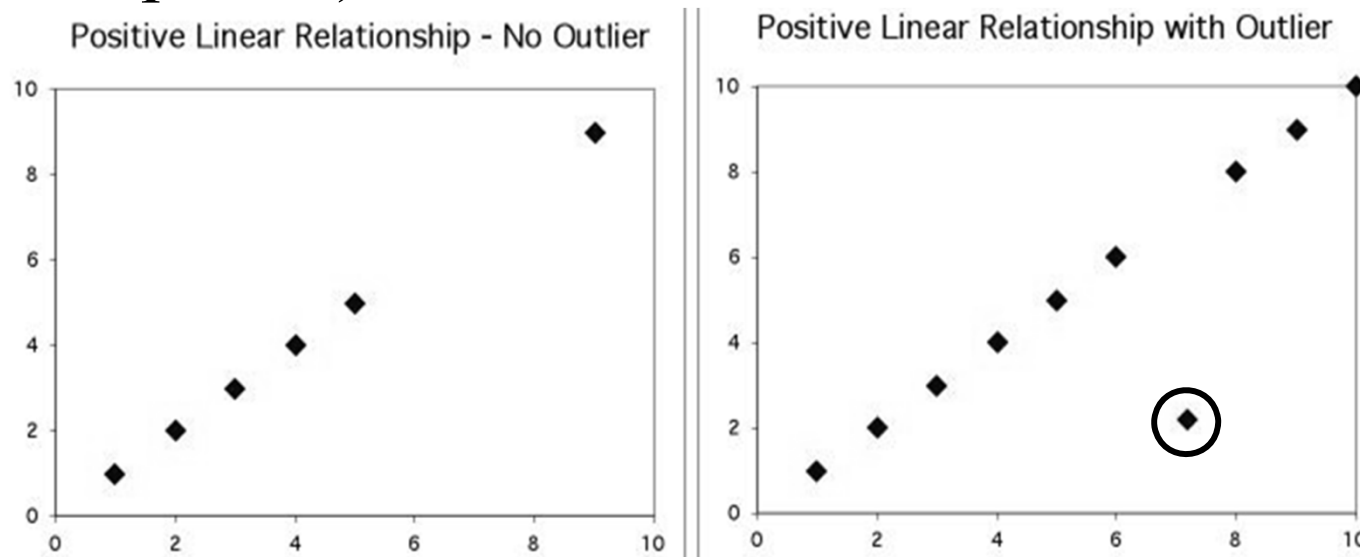


More on Outliers

- An outlier is a data point that comes from a distribution different (in location, scale, or distributional form) from the bulk of the data
- In the real world, outliers have a range of causes, from as simple as
 - operator blunders
 - equipment failures
 - day-to-day effects
 - batch-to-batch differences
 - anomalous input conditions
 - warm-up effects

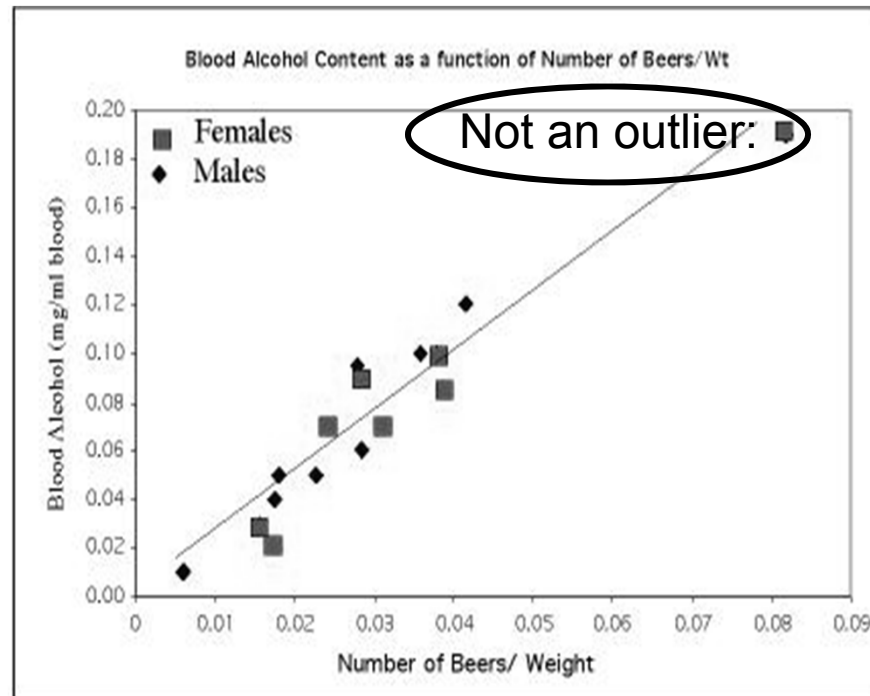
More on Outliers

An **outlier** is a data value that has a very low probability of occurrence (i.e., it is unusual or unexpected).

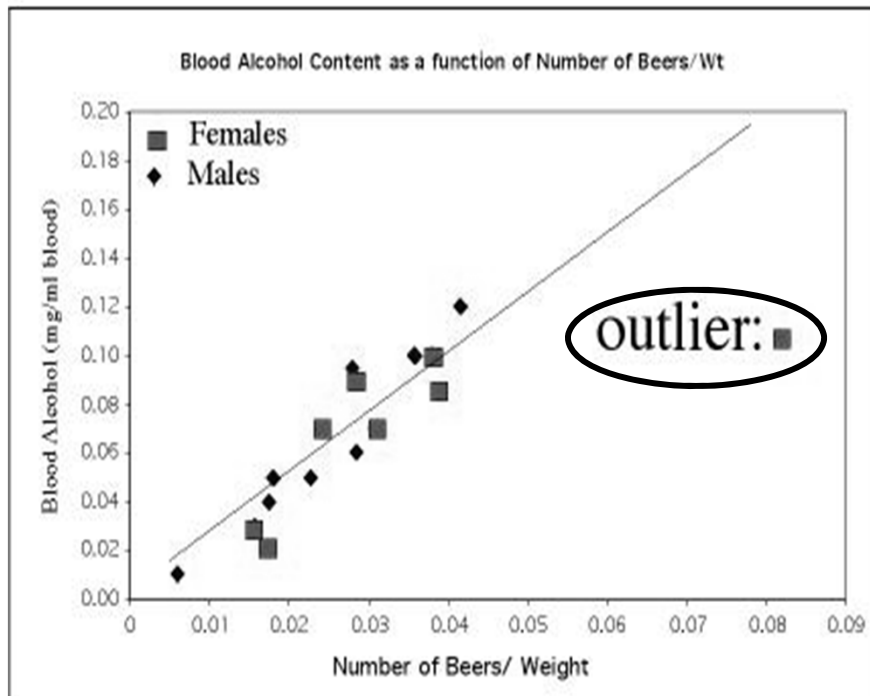


In a scatter plot, outliers are points that fall outside of the overall pattern of the relationship.

Outliers



- The upper right-hand point here is not an outlier of the relationship
- It is what you would expect for this many beers given the linear relationship between beers/weight and blood alcohol.



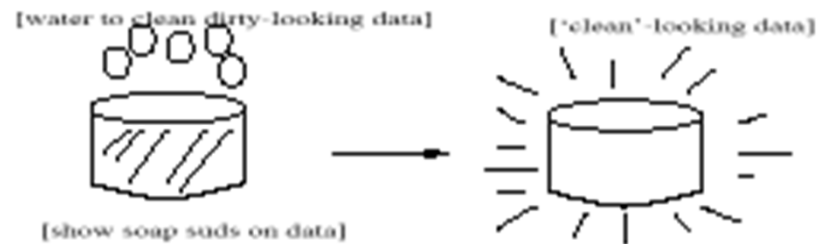
- This point is not in line with the others, so it is an outlier of the relationship.

Major Tasks in Data Preprocessing

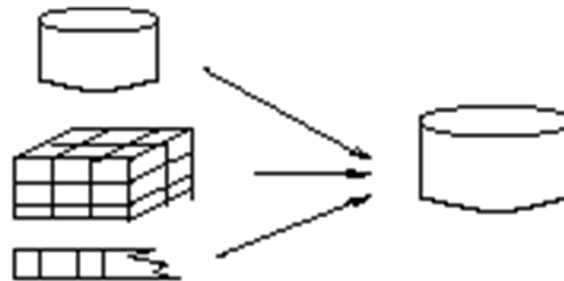
- Data cleaning
 - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
 - Integration of multiple databases, data cubes, or files
- Data transformation
 - Normalization and aggregation
- Data reduction
 - Obtains reduced representation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

Forms of data preprocessing

Data Cleaning



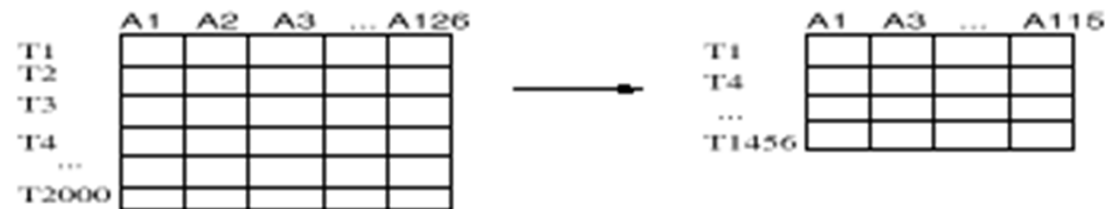
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Data Cleaning

- Data cleaning tasks
 - Fill in missing values
 - Identify outliers and smooth out noisy data
 - Correct inconsistent data

Recover Missing Values

Moving Average

- A **simple moving average** is the unweighted mean of the previous n data points in the time series
- A **weighted moving average** is a weighted mean of the previous n data points in the time series
 - A weighted moving average is more responsive to recent movements than a simple moving average

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- 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

- min-max normalization

$$v' = \frac{v - \textit{min}}{\textit{max} - \textit{min}} (\textit{new_max} - \textit{new_min}) + \textit{new_min}$$

- z-score normalization

$$v' = \frac{v - \textit{mean}}{\textit{stand_dev}}$$

Normalization - Examples

- Suppose that the minimum and maximum values for attribute income are 12,000 and 98,000 respectively. How to map an income value of 73,600 to the range of $[0.0, 1.0]$?
- Suppose that the mean and standard deviation for the attribute income are 54,000 and 15,000. How to map an income value of 73,600 using z-score normalization?

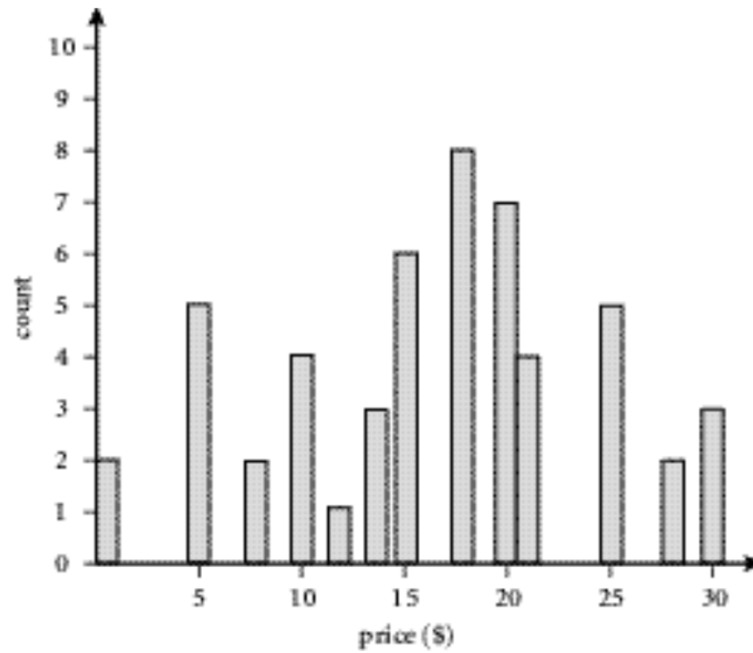
Data Reduction Strategies

- Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
 - Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Data reduction strategies
 - Dimensionality reduction

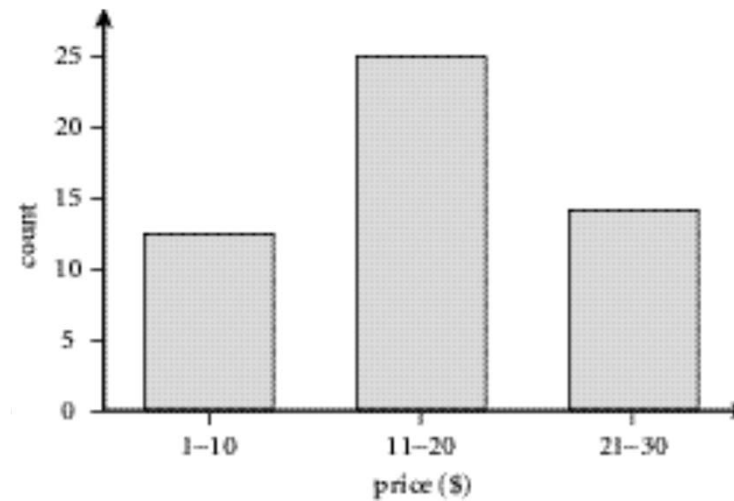
Dimensionality Reduction

- Feature selection (i.e., attribute subset selection):
 - Select a minimum set of features useful for data mining
 - reduce # of patterns in the patterns, easier to understand

Histograms



Singleton buckets



Buckets denoting a
Continuous range of values

Histograms

- How are buckets determined and the attribute values partitioned?
 - Equiwidth: The width of each bucket range is uniform
 - Equidepth: The buckets are created so that, roughly, the frequency of each bucket is constant

Histogram Examples

- Suppose that the values for the attribute *age*:
13, 15, 16, 16, 19, 20, 20, 21, 21, 22, 25, 25, 25, 25, 30, 30, 30,
30, 32, 33, 33, 37, 40, 40, 40, 42, 42

Equiwidth Histogram:

Bucket range	Frequency
13-22	10
23-32	9
33-42	8

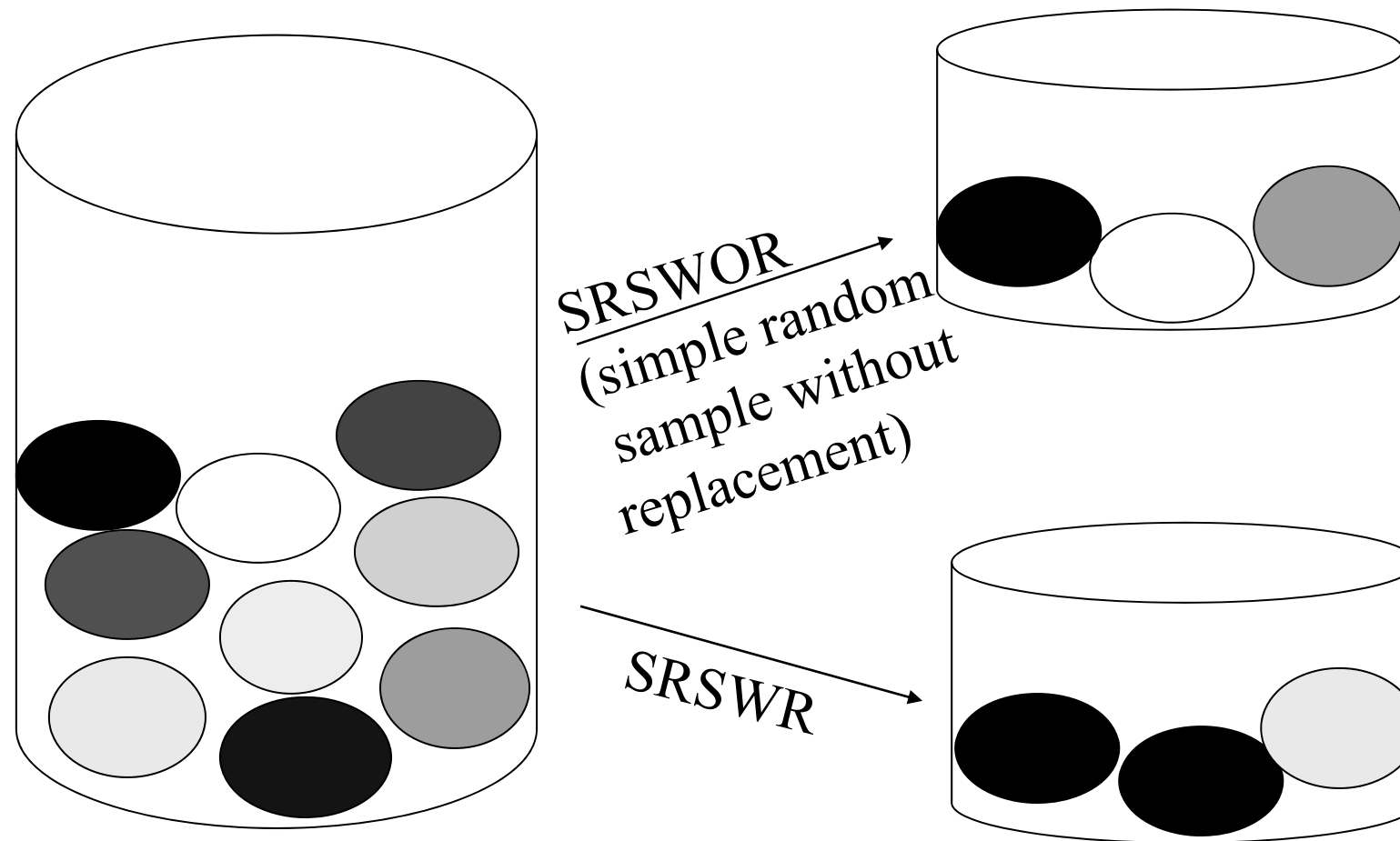
Equidepth Histogram:

Bucket range	Frequency
13-21	9
22-30	9
32-42	9

Sampling

- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a representative subset of the data
 - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
 - Stratified sampling:
 - Approximate the percentage of each class (or subpopulation of interest) in the overall database
 - Used in conjunction with skewed data
- Sampling may not reduce database I/Os (page at a time).

Sampling



ECLT 5810 Raw Data

Data Preprocessing

