

Data Preprocessing

- Lecture 3: Overview of data preprocessing/ Descriptive data summarization
- Lecture 4: Data cleaning / Data integration/ transformation
- Lecture 5: Data reduction /Discretization and concept hierarchy generation and summary

Lecture 3

Why Data Preprocessing?

- **Data in the real world is dirty**
 - **incomplete**: lacking attribute values, lacking certain attributes of interest, e.g., occupation=" "
 - **noisy**: containing errors or outliers
 - e.g., Salary="-10"
 - **inconsistent**: containing discrepancies in codes or names
 - e.g., Age="42" Birthday="03/07/1999"
 - e.g., Was rating "1,2,3", now rating "A, B, C"
 - e.g., discrepancy between duplicate records

Lecture 3

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Lecture 3

Why Is Data Dirty?

- Incomplete data may come from
 - "Not applicable" data value when collected
 - Different considerations between the time when the data was collected and when it is analyzed.
 - Human/hardware/software problems
- Noisy data (incorrect values) may come from
 - Faulty data collection instruments
 - Human or computer error at data entry
 - Errors in data transmission
- Inconsistent data may come from
 - Different data sources
 - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Lecture 3

1

3

2

4

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
 - Quality decisions must be based on quality data
 - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
 - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

Lecture 3

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

Lecture 3

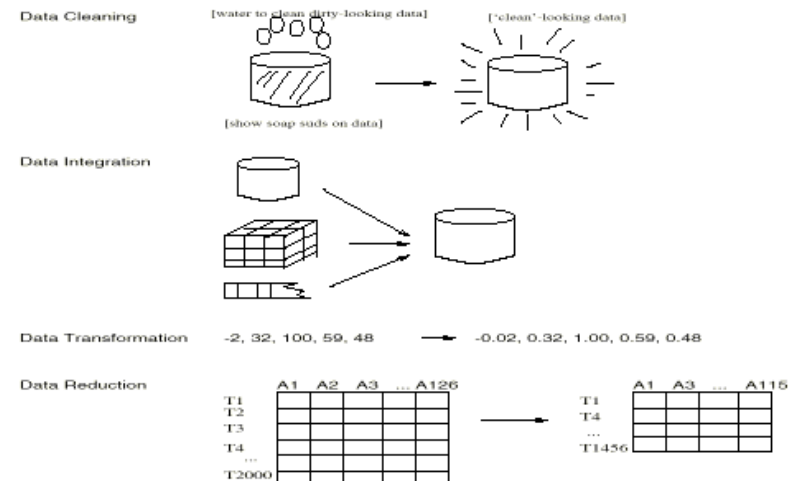
Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
 - Accuracy
 - Completeness
 - Consistency
 - Timeliness
 - Believability
 - Value added
 - Interpretability
 - Accessibility
- Broad categories:
 - Intrinsic, contextual, representational, and accessibility

5

Lecture 3

Forms of Data Preprocessing



7

Lecture 3

6

8

Mining Data Descriptive Characteristics

- **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

Lecture 3

Measuring the Central Tendency

- **Mean (sample vs. population):** $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
 - 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:** A holistic measure
 - Middle value if odd number of values, or average of the middle two values otherwise
 - Estimated by interpolation (for *grouped data*): $median = L_1 + \left(\frac{\frac{n}{2} - (\sum f)_l}{f_{median}} \right) c$
- **Mode**
 - Value that occurs most frequently in the data
 - Unimodal, bimodal, trimodal
 - Empirical formula: $mean - mode = \gamma \times (mean - median)$

Lecture 3

11

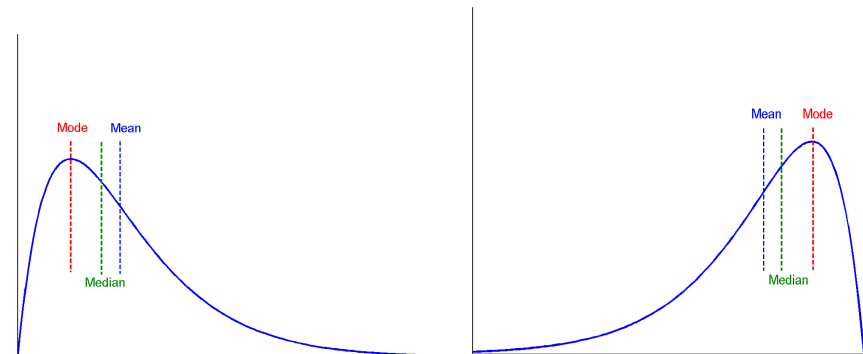
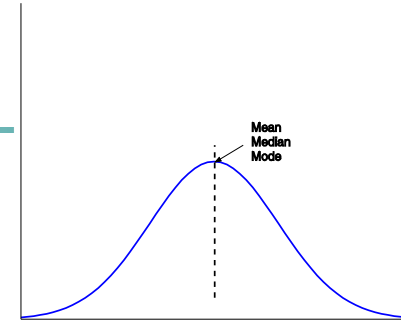
Measuring Data Central Tendency

- **Distributive Measure:** compute the measure from each subset and merge the results (sum, count)
- **Algebraic Measure:** apply an algebraic function to one or more distributive measures (average=sum()/count())
- **Holistic Measure:** be computed based on the entire data set (median)

Lecture 3

Symmetric vs. Skewed Data

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



10

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, whiskers, and plot outlier individually
 - **Outlier:** usually, a value higher/lower than $1.5 \times IQR$
- **Variance and standard deviation (sample: s , population: σ)**

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

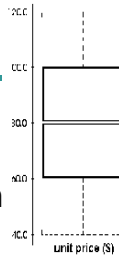
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

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

Lecture 3

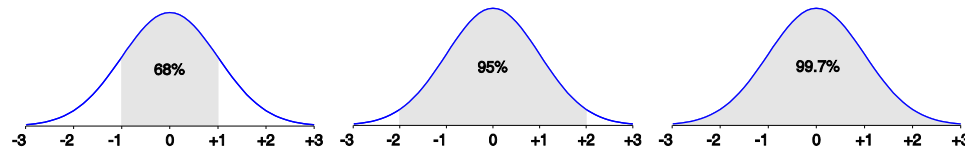
Boxplot Analysis

- **Five-number summary** of a distribution:
Minimum, Q_1 , Median, Q_3 , 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 IRQ
 - The median is marked by a line within the box
 - Whiskers: two lines outside the box extend to Minimum and Maximum



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

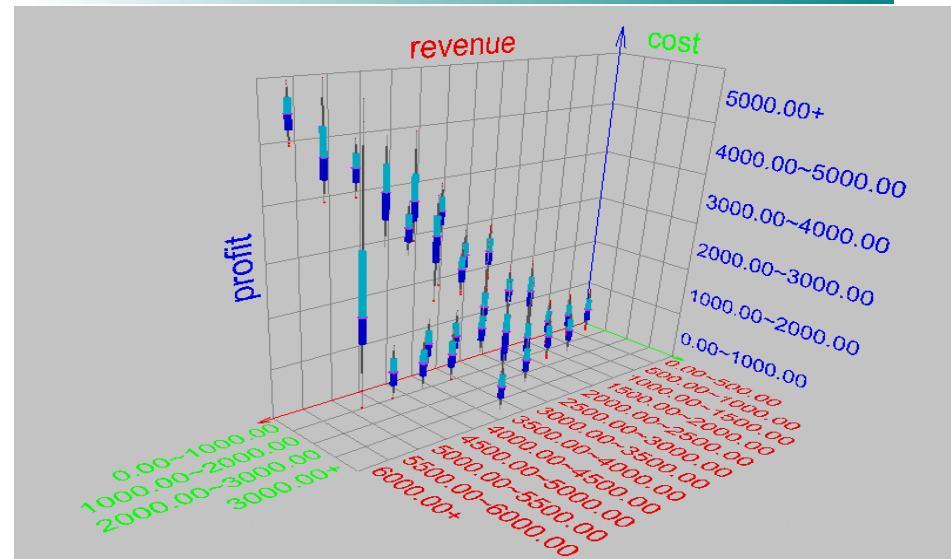


13

Lecture 3

14

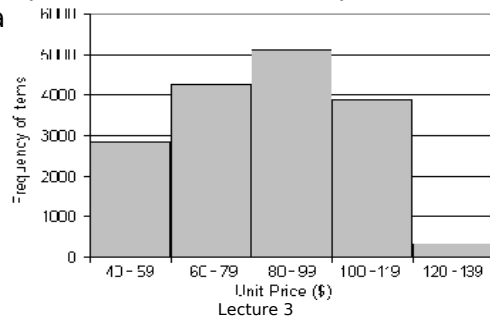
Visualization of Data Dispersion: Boxplot Analysis



15

Histogram Analysis

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



Graphic Displays of Basic Statistical Descriptions

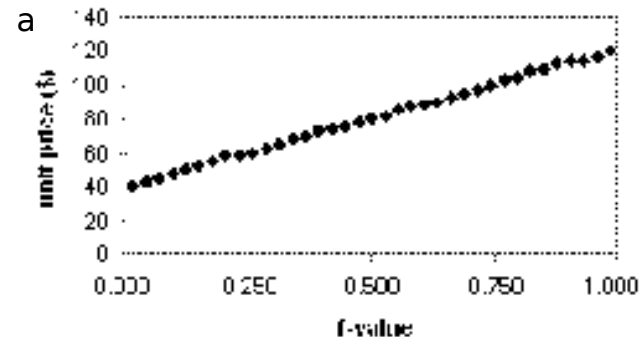
- Histogram: (shown before)
- Boxplot: (covered before)
- Quantile plot: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- Quantile-quantile (q-q) plot: graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane
- Loess (local regression) curve: add a smooth curve to a scatter plot to provide better perception of the pattern of dependence

Lecture 3

17

Quantile Plot

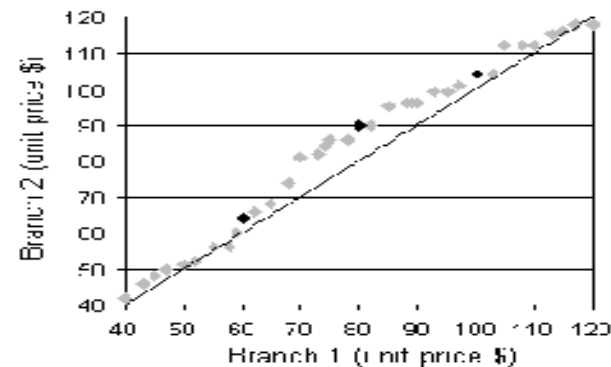
- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
 - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i % of the data



18

Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another

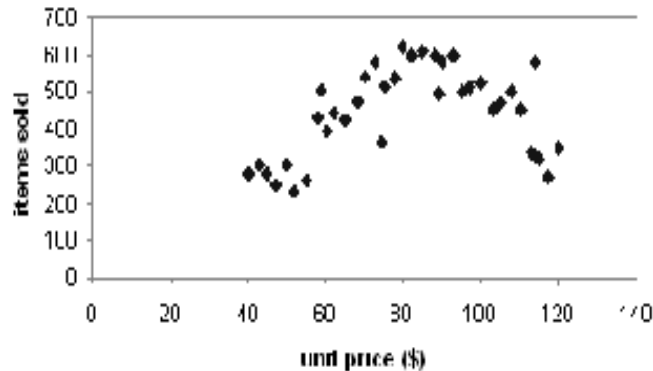


19

20

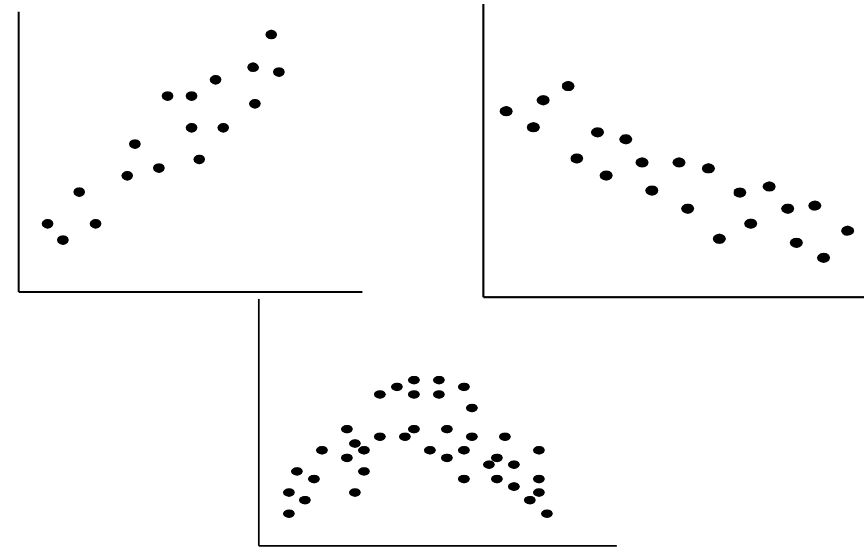
Scatter plot

- 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



21

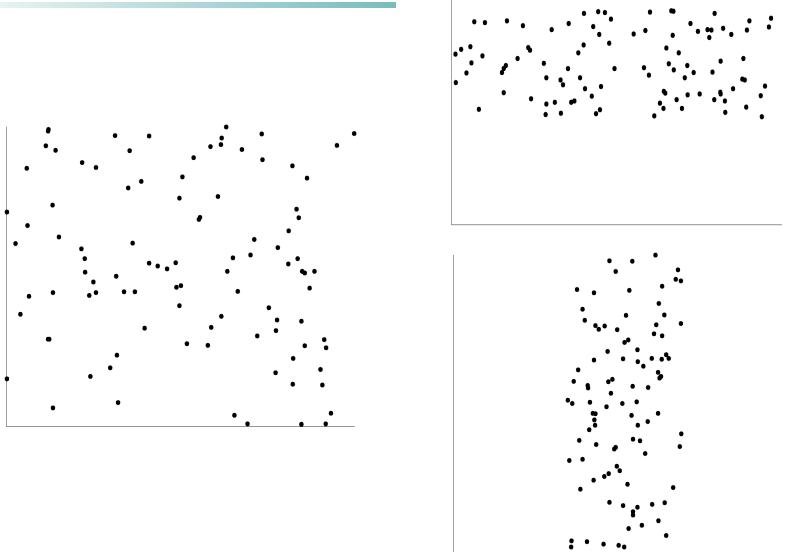
Positively and Negatively Correlated Data



Lecture 3

22

Not Correlated Data

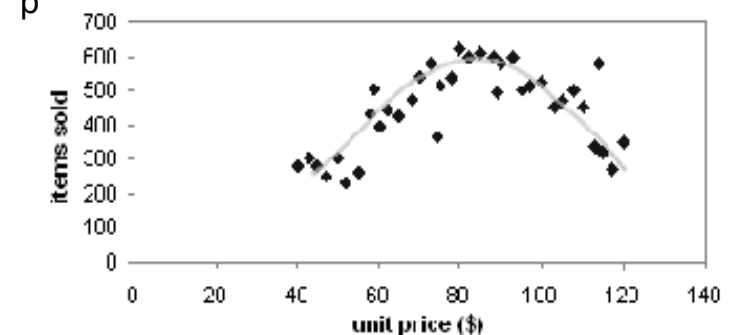


Lecture

3

Loess Curve

- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the p



24