Data Mining:

Chapter 3

Pre-mining

Outline

- Preliminaries
  - Data objects and Attribute Types
  - Basic Statistics
  - Graphic Displays
- Pre-mining / Data Preprocessing
  - Data Preprocessing: An Overview
  - Data Cleaning

Data Mining:

Concepts and Techniques

(3rd ed.)

— Chapter 2 —

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

Types of Datasets

- Record data
  - Relational records: Relational table, highly structured
  - Data matrix: Numerical matrix, cross tabs
  - Transaction data
  - Document data: Term-frequency vector of documents
- Graphs and networks
  - Transportation network
  - World Wide Web
  - Molecular structures
  - Social or information networks

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

no relation

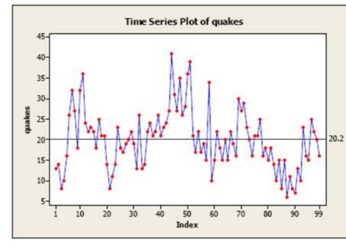
Car:

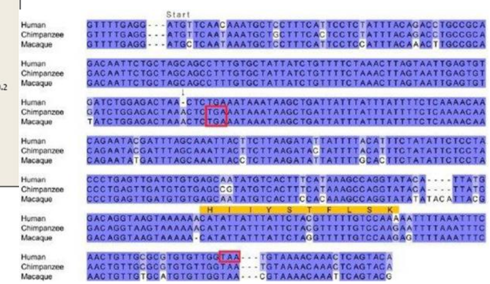
Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

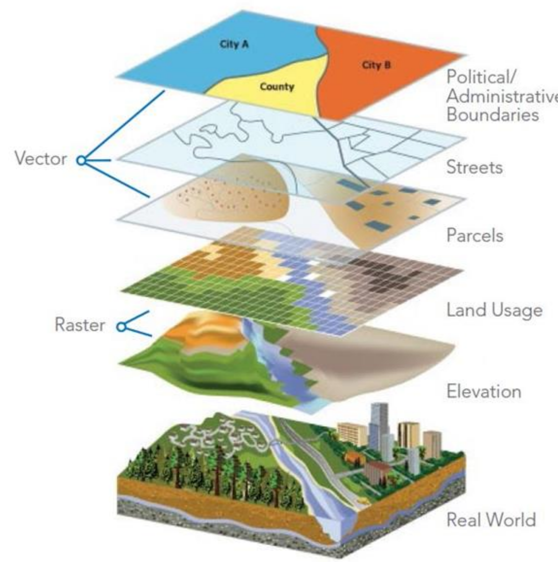
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Types of Datasets

- Ordered data
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential data: transaction sequence
  - Genetic sequence data
- Spatial / visual / multimedia data
  - Spatial data: maps
  - Image data
  - Video data

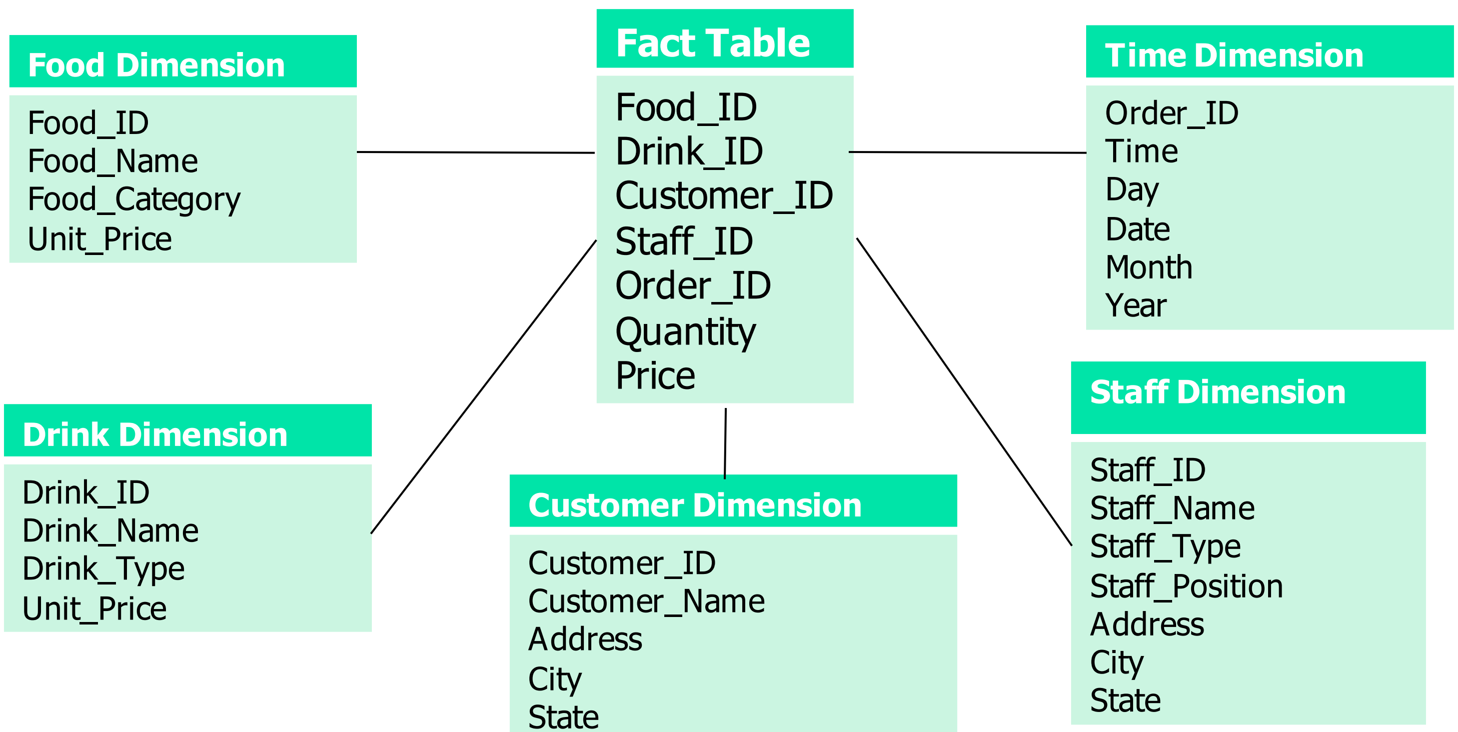






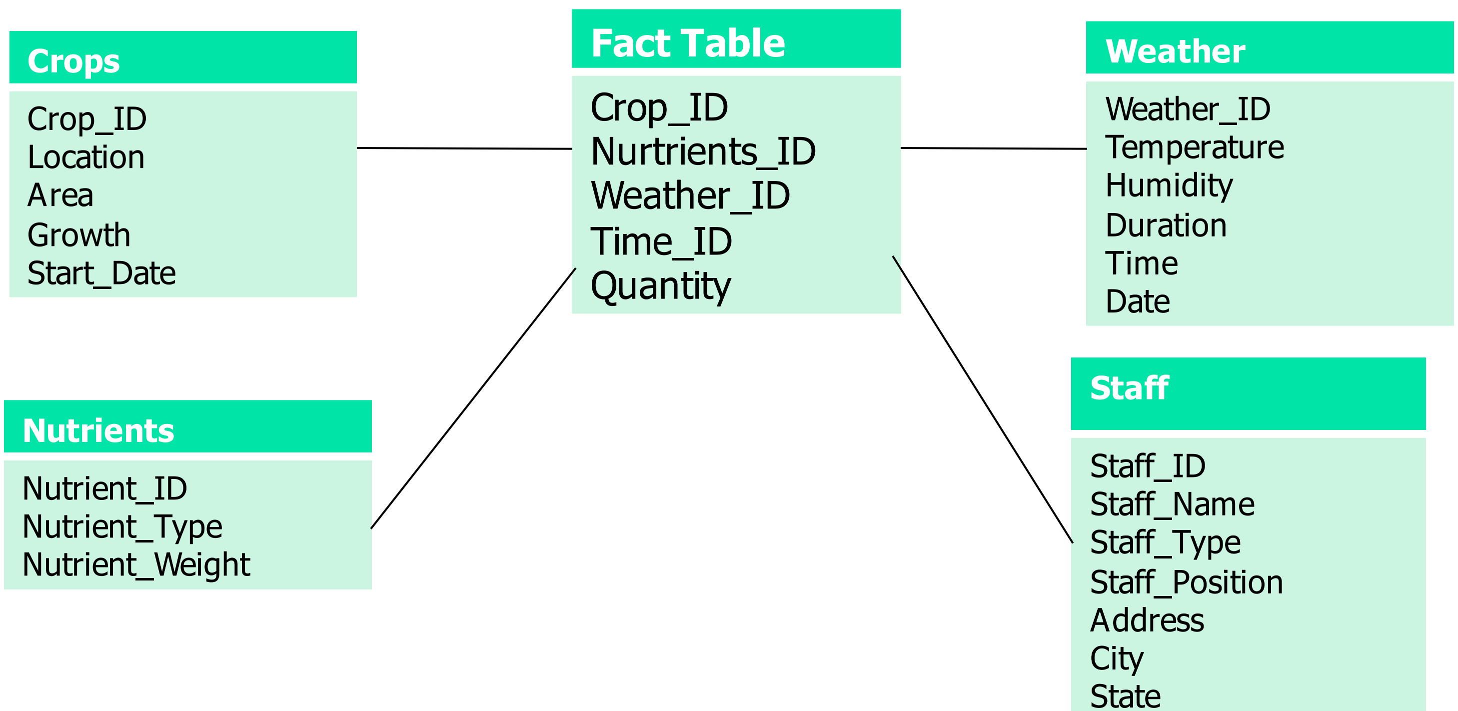
Revisit

- A restaurant owner wants to identify common set of preferences among customers.
  - Sales data
  - Food, Drink, Customer, Staff, Order



Revisit

- Farmer wants to predict yield of rice in the next harvest
  - Production data
  - Crops, Nutrients, Weather



Important Characteristics

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Small amounts of appearance spread over large area
- Resolution
  - Patterns depending on scale / scope
- Distribution
  - Centrality and dispersion statistics

Data Objects

- Data sets are made up of data objects
- Data object represents an entity
  - Sales database: customer, store items, sales
  - Medical database: patients, treatments
  - University database: students, professors, courses
- Also called *samples, examples, instances, tuples, data points, records*
- **Data objects** are described by **attributes**
- Database: rows of data objects, columns of attributes

Data Attributes

- Attributes are also called *dimensions, features, variables*
  - A data field representing a characteristic or feature of a data object
  - E.g. customer_ID, name, address
- Attributes are distinguishable by **types**
  - Nominal: Categories
  - Binary: Only 2 categories
  - Ordinal: Ordered / Ranked
  - Numeric: Quantitative
    - Interval-scaled or Ratio scaled

Attribute Types

- Nominal: categories, states, or "name of things"
  - Hair_color = {auburn, black, blond}
  - Marital status, occupation
- Binary: nominal attribute with only 2 states
  - Symmetrical binary: both outcomes equally important, e.g. gender
  - Asymmetrical binary: outcomes not equally important, e.g. medical test
- Ordinal: values have meaningful order but unknown magnitude
  - Size = {small, medium, large}

Numerical Attribute Types

- Attributes measured in quantities: integer or real-valued
- Interval-scaled
  - No true zero-point
    - E.g. 0 Celsius / 0 Fahrenheit does not indicate "no temperature"
  - Values have order
  - Differences between values can be quantified
    - 20 Celsius is 5 degrees higher than 15 Celsius
    - The year 2022 and 2012 are 10 years apart

Numerical Attribute Types

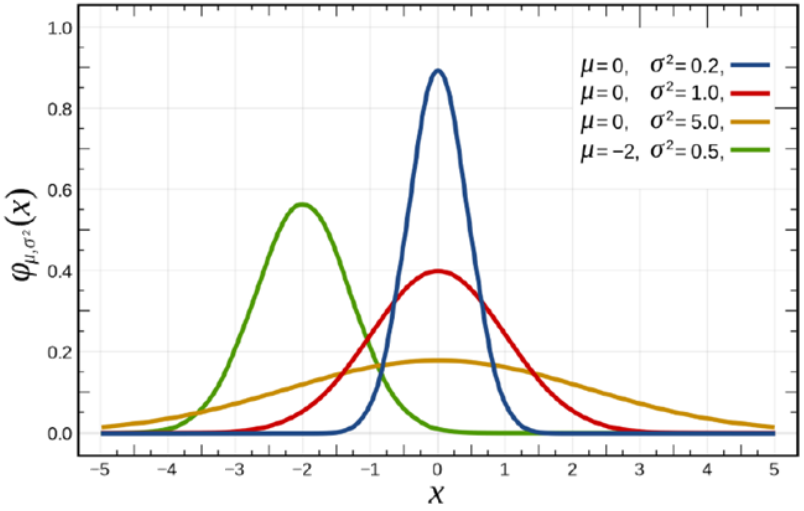
- Ratio-scaled
  - Has inherent zero-point
    - 0 Kelvin (-273.15 Celsius) is considered a true zero-point because it indicates zero kinetic energy
  - Values can indicate orders of magnitude higher/lower in the unit of measurement
    - 10 Kelvin is twice as high as 5 Kelvin
  - E.g. Temperature in Kelvin, length, counts, monetary quantities

Discrete vs. Continuous

- Attributes can also be distinguishable according to the value types
- Discrete attributes
  - Has only a finite or countably infinite set of values
    - e.g. zip codes, profession, set of words in collection of documents, number of cars
  - Can be both categorical or numerical
  - Binary is a special case of this
- Continuous attributes
  - Has real numbers as attribute values
    - e.g. temperature, height, or weight
  - Typically represented as floating-point variables
  - Practically, real values can only be measured and represented using finite number of digits (discrete)

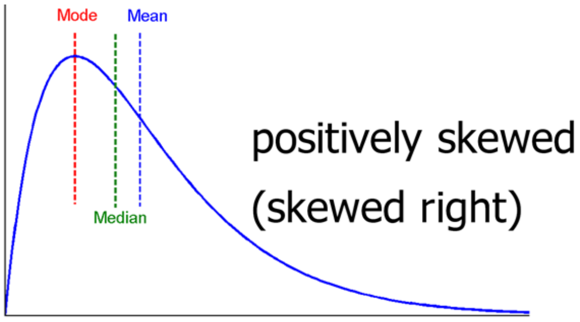
Basic Statistical Description

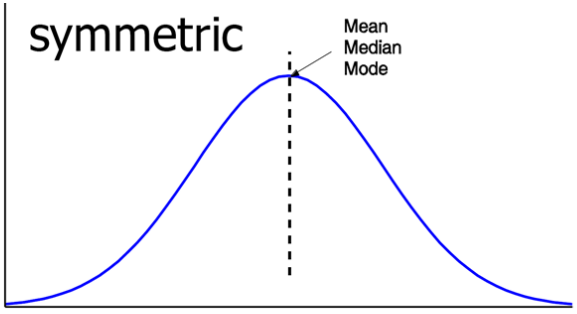
- Prior to data mining, it is necessary to explore and understand data
- Typical explorations
  - Central tendency: mean, median, mode, skew
  - Variation: Variance, outliers
  - Spread: standard deviation, range, kurtosis

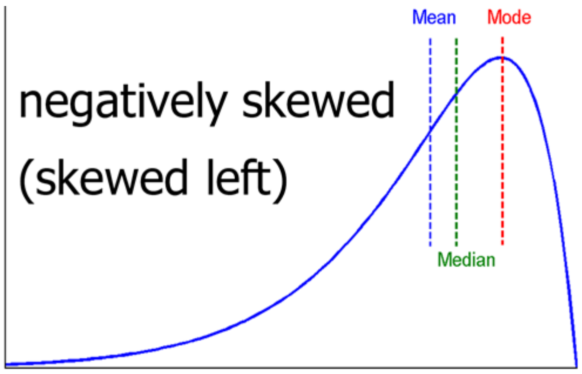


Central Tendency

- Mean: average value of data
- Median: center value of data
- Mode: most frequent value of data
- Skew: shape of distribution

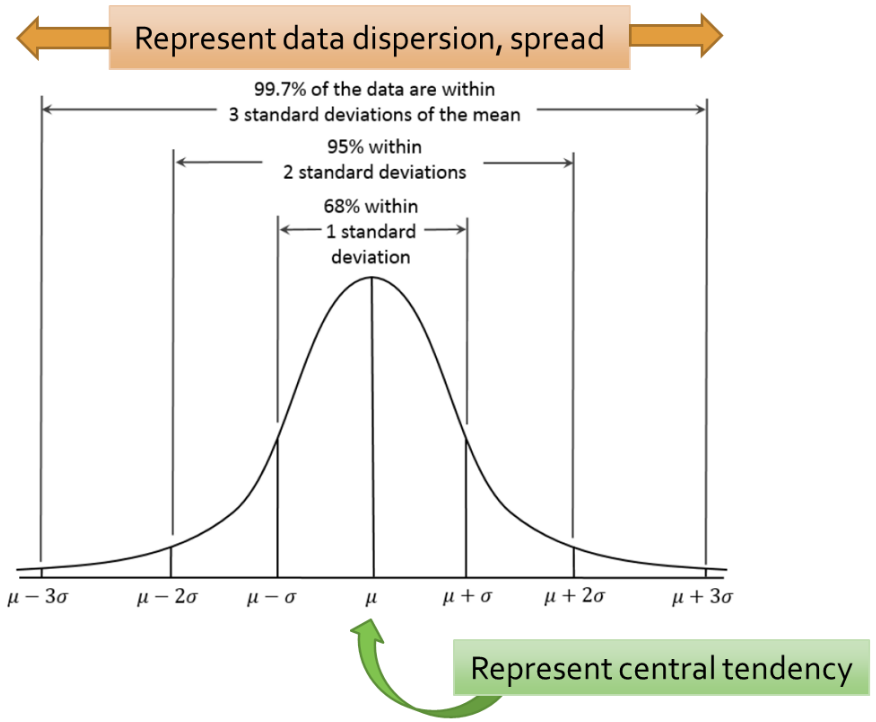






Distribution

- Mean, variance, and standard deviation are main indicators
- Standard deviation: measure of spread
  - Standard to know "normal" from "extra large" or "extra small"
- Z-score: number of standard deviations away from mean

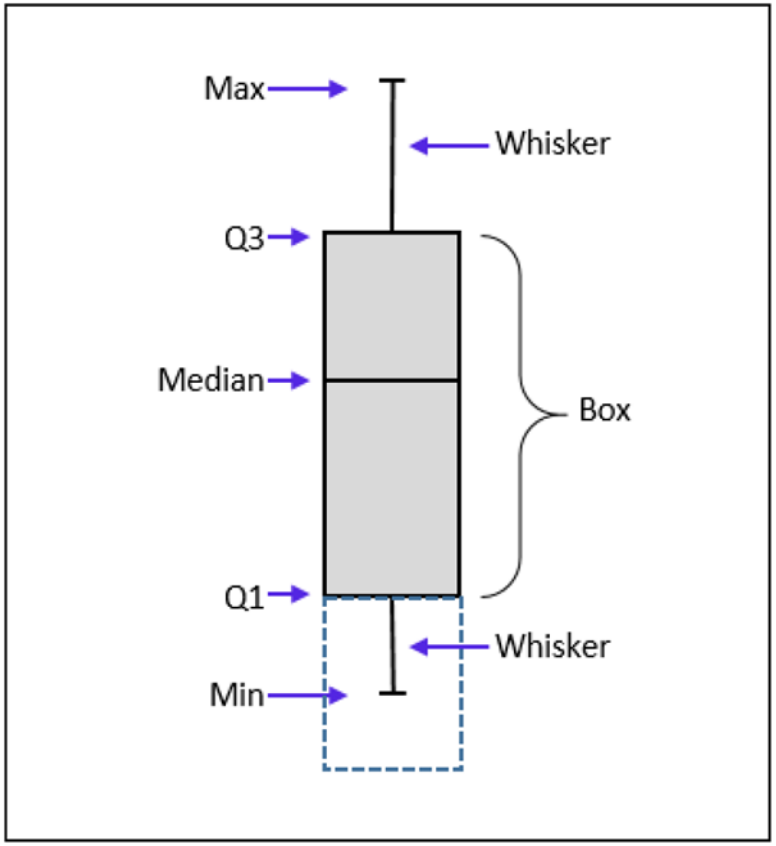


Graphic Displays of Statistics

- Boxplot
  - Five-number summary to show data dispersion
- Histogram
  - Bars to show frequency of appearance of data points in specified ranges
- Quantile-quantile (q-q) plot
  - Graphs quantile of one distribution against quantile of another distribution for comparison
- Scatter plot
  - Pairs of values used as coordinates to plot points in a plane

Quartiles and Boxplots

- 5-number summary:
  - Min, Q1, Median, Q3, Max
- Can indicate outliers using Inter-quartile range,
  - IQR = |Q1 – Q3|
  - Upper outliers > Q3+1.5IQR
  - Lower outliers < Q1-1.5IQR
- Outliers: points beyond specified threshold

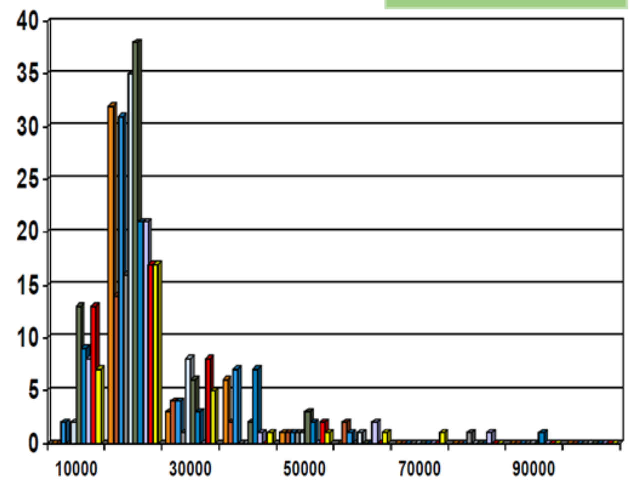


Histogram Analysis

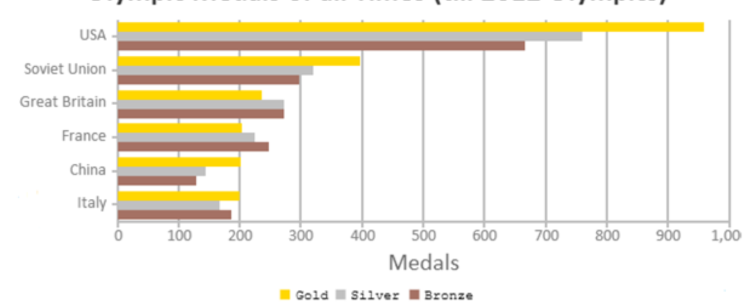
- Graphical display of tabulated frequencies shown in bars

Histogram	Bar Charts
Show distribution of variables	Used to compare variables
Plot binned quantitative data	Plot categorical data
Set order of bins	Bars can be reordered

Histogram



Olympic Medals of all Times (till 2012 Olympics)

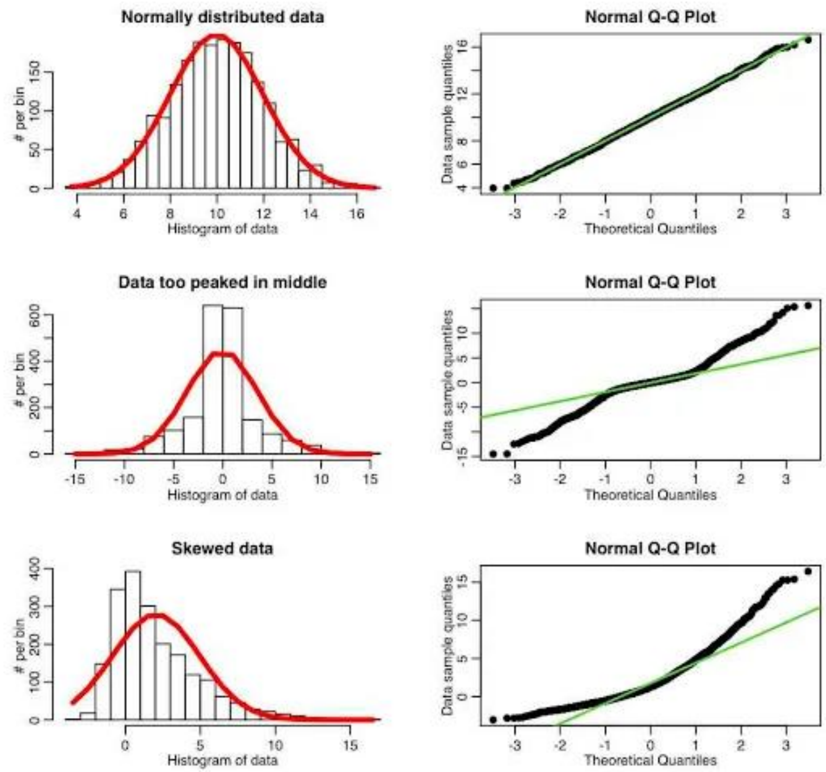


Bar chart

- Histogram often tells more than Boxplots

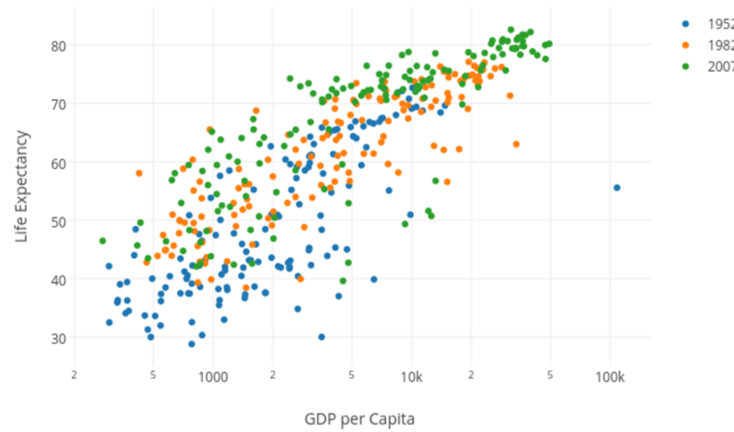
QQ Plot

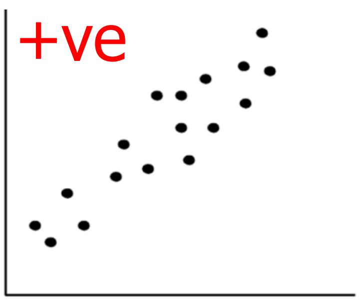
- Used to find the type of distribution for a random variable (e.g. normal, exponential, etc.)
- Theoretical quantiles (a normal distribution with mean=0 and standard deviation=1) on the x-axis **vs.** the ordered values of data on the y-axis
  - Smooth straight line if Gaussian
- Can be used to compare distributions of 2 sets of data by plotting the quantiles (one axis for each set of data)

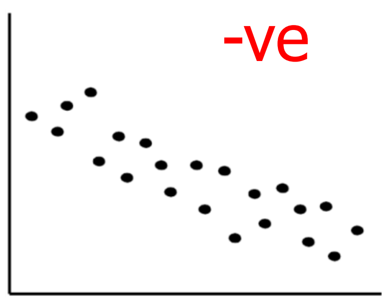


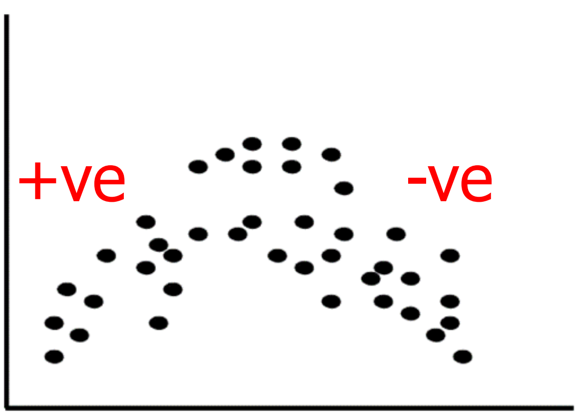
Scatter Plot

- Used to observe data patterns of 2 variables
  - Extendable to more with more plots of multidimensional plot
- Can identify clusters, outliers, etc.

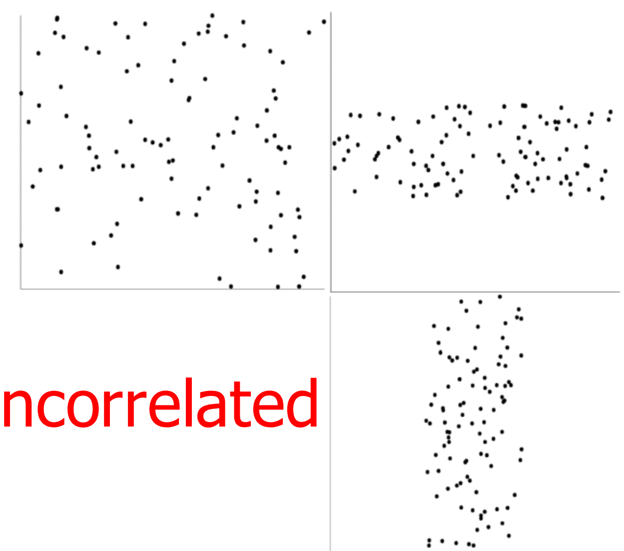








Uncorrelated



Outline

- Preliminaries
  - Data objects and Attribute Types
  - Basic Statistics
  - Graphic Displays
- Pre-mining / Data Preprocessing
  - Data Preprocessing: An Overview
  - Data Cleaning

Data Mining:

Concepts and Techniques

(3rd ed.)

— Chapter 3 —

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

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

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 reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=" " (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary*="−10" (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*="42", *Birthday*="03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- 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
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

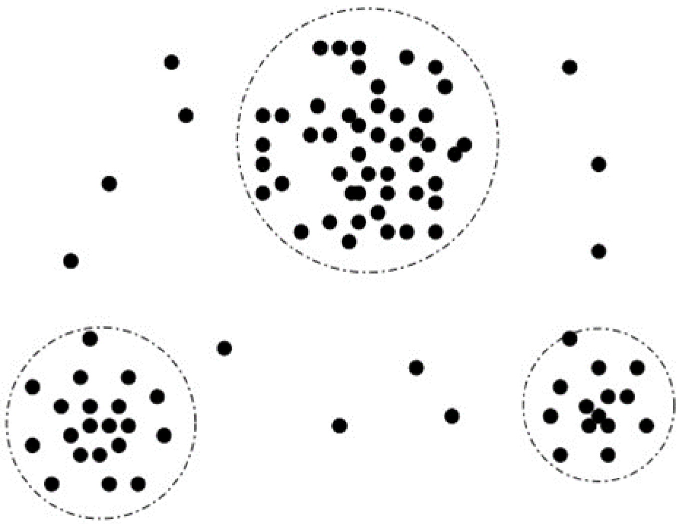
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

How to Handle Noisy Data?

- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection (semi-supervised)
  - detect suspicious values and check by human (e.g., deal with possible outliers)



Data Cleaning as a Process

- Data discrepancy detection
  - Detect lack of compatibility between two or more facts that should be similar
    - i.e. caused by poorly designed form
  - How to handle?
    - Use metadata (e.g., domain, range, dependency, distribution)
    - Ask questions like:
      - Do all values fall within expected range?
      - Are there any known dependencies between attributes?
      - Are values more than 2 standard deviations away from the mean flagged as potential outliers?
    - Check inconsistent data representations
      - "2010/12/25" and "25/12/2010" for date
    - Check field overloading
      - developers squeeze new attribute definitions into unused (bit) portions of already defined attributes

Data Cleaning as a Process

- Data discrepancy detection
  - Check uniqueness rule
    - each value of certain given attributes must be unique
  - Check consecutive rule
    - no missing values between the lowest and highest values for the attribute, and that all values must also be unique
  - Check null rule
    - the use of blanks, question marks, special characters, or other strings that may indicate a value for a given attribute is not available
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)

Data Cleaning as a Process

- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Discrepancy detection and data transformation
  - Iterative and interactive (e.g., Potter's Wheels)
- Example of opensource tool for data cleaning:
  - http://openrefine.org (formally Google refine)

Summary

- **Datasets**: Structured, unstructured, record, graphical, ordered, spatial
- **Data**: Data objects, attributes, nominal, binary, ordinal, numerical
- **Statistics**: Mean, median, mode, skew, variance, standard deviation, z-score
- **Graphics**: Boxplot, histogram, qq plot, scatter plot
- **Data quality**: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values, outliers

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Comm. of ACM, 42:73-78, 1999
- A. Bruce, D. Donoho, and H.-Y. Gao. Wavelet analysis. *IEEE Spectrum*, Oct 1996
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- J. Devore and R. Peck. *Statistics: The Exploration and Analysis of Data*. Duxbury Press, 1997.
- H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita. Declarative data cleaning: Language, model, and algorithms. *VLDB'01*
- M. Hua and J. Pei. Cleaning disguised missing data: A heuristic approach. *KDD'07*
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), Dec. 1997
- H. Liu and H. Motoda (eds.). *Feature Extraction, Construction, and Selection: A Data Mining Perspective*. Kluwer Academic, 1998
- J. E. Olson. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 2003
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- T. Redman. *Data Quality: The Field Guide*. Digital Press (Elsevier), 2001
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995