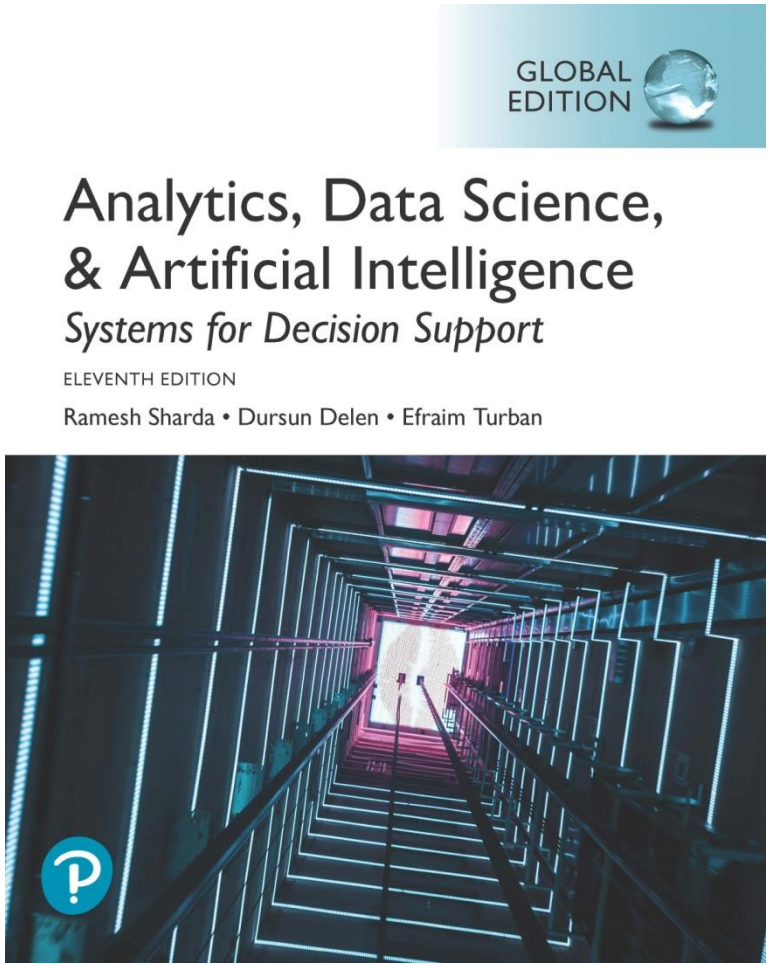
Analytics, Data Science and AI: Systems for Decision Support

Eleventh Edition, Global Edition



Chapter 3

Nature of Data, Statistical Modeling and Visualization

Learning Objectives

3.1 Understand the nature of data as it relates to business intelligence (BI) and analytics

3.2 Describe statistical modeling and its relationship to business analytics

3.3 Learn about descriptive and inferential statistics

3.4 Understand the importance of data/information visualization

3.5 Learn different types of visualization techniques

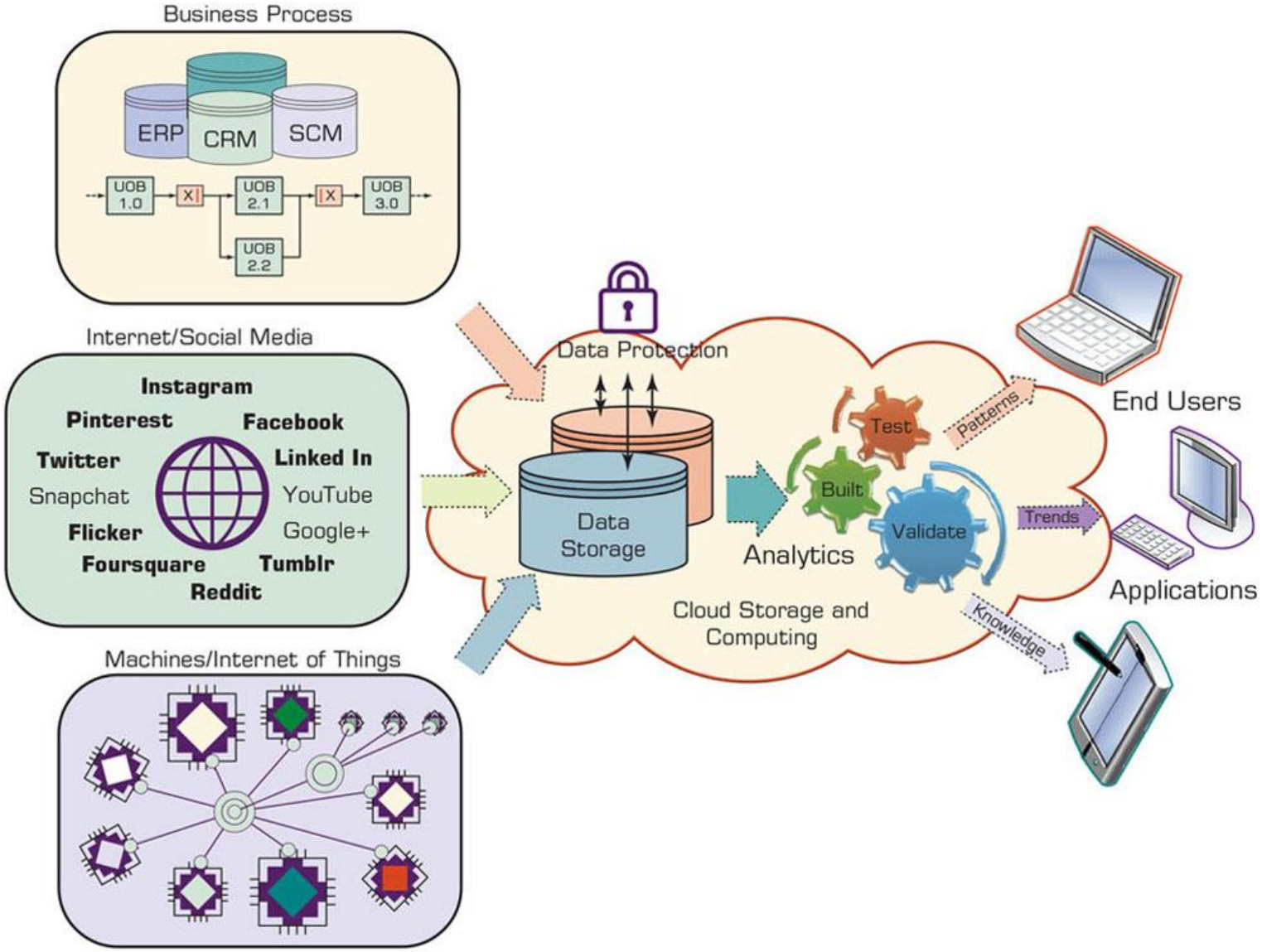
3.6 Appreciate the value that visual analytics brings to business analytics

3.7 Know the capabilities and limitations of dashboards

The Nature of Data (1 of 2)

- **Data**: a collection of facts
  - usually obtained as the result of experiences, observations, or experiments
- Data may consist of numbers, words, images, …
- Data is the lowest level of abstraction (from which information and knowledge are derived)
- Data is the source for information and knowledge
- Data quality and data integrity → critical to analytics
  - With data coming from several sources, maintain data quality and integrity is a challenge.

The Nature of Data (2 of 2)



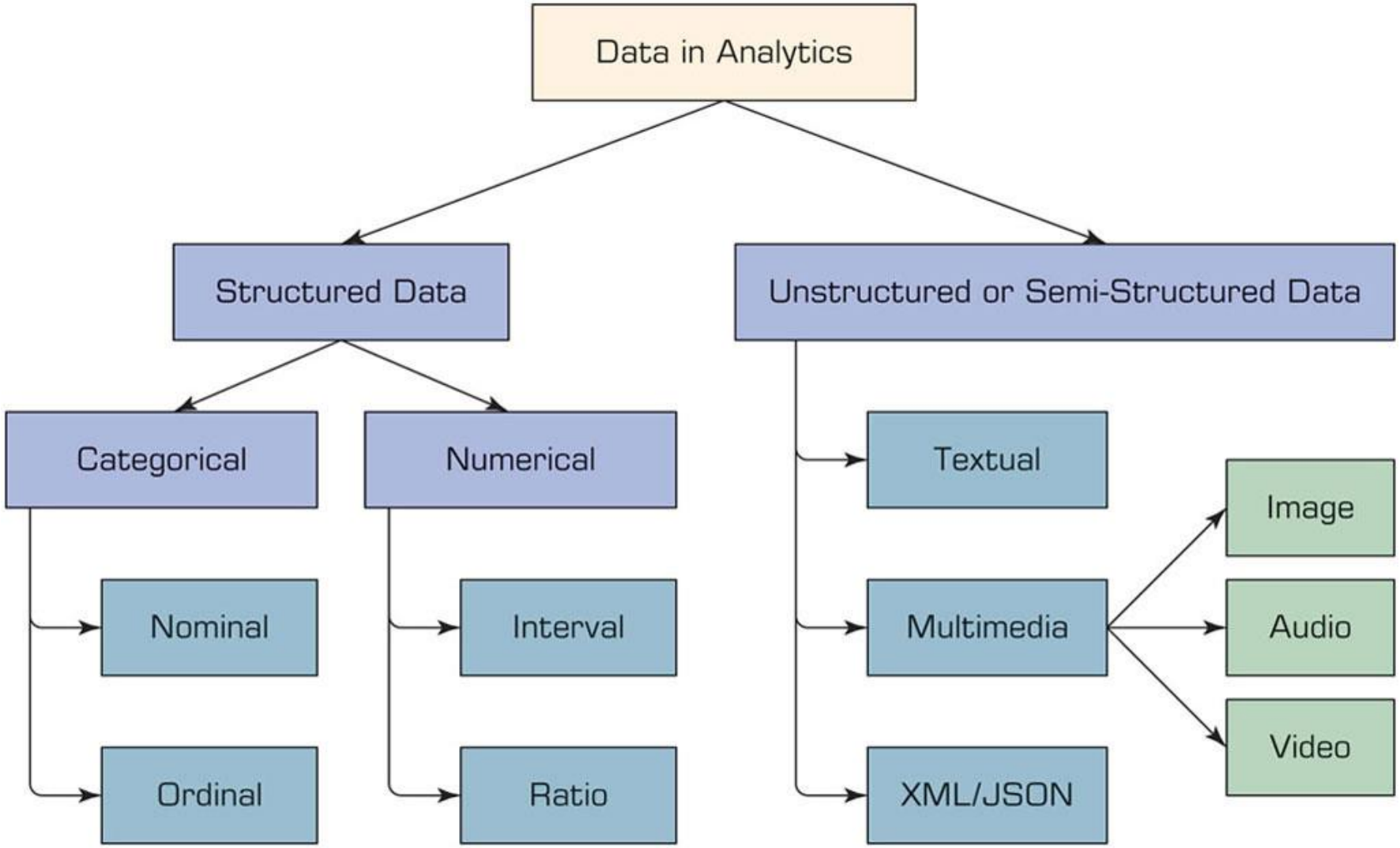
Metrics for Analytics ready Data

- Data source reliability → do you have confidence in the data source?
- Data content accuracy → do you have the right data for the job?
- Data accessibility → can we easily get to the data when we need to?
- Data security and data privacy → CRUD!
- Data richness → data availability & completeness.
- Data consistency → integration & merging.
- Data currency/data timeliness → up-do-date
- Data granularity → level of detail <> aggregate.
- Data validity and data relevancy → data definition; analysis contamination.

A Simple Taxonomy of Data (1 of 2)

- Data (datum—singular form of data): facts
- Structured data
  - Targeted for computers to process
  - Numeric versus nominal
- Unstructured/textual data
  - Targeted for humans to process/digest
- Semi-structured data?
  - XML, HTML, Log files, etc.
- Data taxonomy…

A Simple Taxonomy of Data (2 of 2)



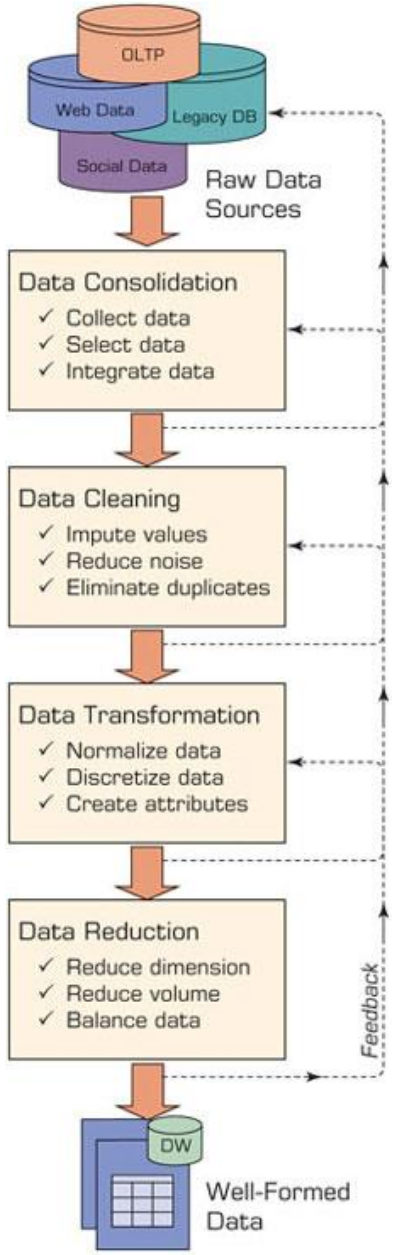
The Art and Science of Data Preprocessing (1 of 2)

- The real-world data is dirty, misaligned, overly complex, and inaccurate
  - Not ready for analytics!
- Readying the data for analytics is needed
  - Data preprocessing
    - Data consolidation
    - Data cleaning
    - Data transformation
    - Data reduction
- Art – it develops and improves with experience

The Art and Science of Data Preprocessing (2 of 2)

- Data reduction

1. Variables
   - Dimensional reduction
   - Variable selection
2. Cases/samples
   - Sampling
   - Balancing / stratification

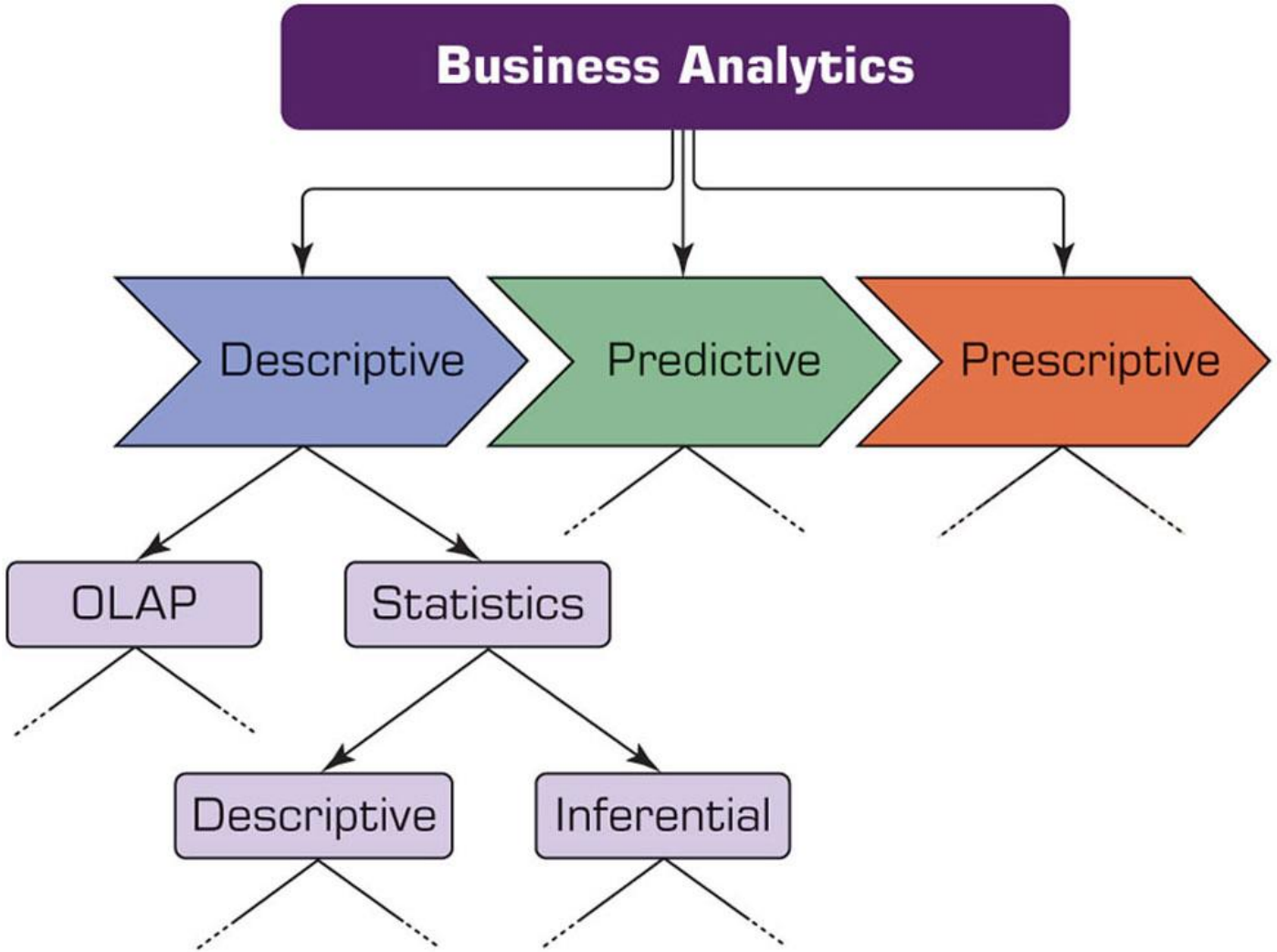


Data Preprocessing Tasks and Methods

Table 3.1 A Summary of Data Preprocessing Tasks and Potential Methods.

Main Task	Subtasks	Popular Methods
Data consolidation	Access and collect the data Select and filter the data Integrate and unify the data	SQL queries, software agents, Web services. Domain expertise, SQL queries, statistical tests. SQL queries, domain expertise, ontology-driven data mapping.
Data cleaning	Handle missing values in the data	Fill in missing values (imputations) with most appropriate values (mean, median, min/max, mode, etc.); recode the missing values with a constant such as “ML”; remove the record of the missing value; do nothing.
	Identify and reduce noise in the data	Identify the outliers in data with simple statistical techniques (such as averages and standard deviations) or with cluster analysis; once identified, either remove the outliers or smooth them by using binning, regression, or simple averages.
	Find and eliminate erroneous data	Identify the erroneous values in data (other than outliers), such as odd values, inconsistent class labels, odd distributions; once identified, use domain expertise to correct the values or remove the records holding the erroneous values.
Data transformation	Normalize the data	Reduce the range of values in each numerically valued variable to a standard range (e.g., 0 to 1 or −1 to +1) by using a variety of normalization or scaling techniques.
	Discretize or aggregate the data	If needed, convert the numeric variables into discrete representations using range- or frequency-based binning techniques; for categorical variables, reduce the number of values by applying proper concept hierarchies.
	Construct new attributes	Derive new and more informative variables from the existing ones using a wide range of mathematical functions (as simple as addition and multiplication or as complex as a hybrid combination of log transformations).
Data reduction	Reduce number of attributes	Use principal component analysis, independent component analysis, chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Perform random sampling, stratified sampling, expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or undersample the more represented classes.

Statistical Modeling for Business Analytics (1 of 2)



Statistical Modeling for Business Analytics (2 of 2)

- Statistics
  - A collection of mathematical techniques to characterize and interpret data
- Descriptive Statistics
  - Describing the data (as it is)
- Inferential statistics
  - Drawing inferences about the population based on a sample data
- Descriptive statistics for descriptive analytics

Descriptive Statistics Measures of Centrality Tendency (1 of 2)

- Arithmetic mean

$$\bar{x} = \frac{x_1 + x_2 + \ \ldots + x_n}{n} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Median
  - The number in the middle
- Mode
  - The most frequent observation

Descriptive Statistics Measures of Dispersion (1 of 2)

- Dispersion
  - Degree of variation in a given variable
- Range
  - Max - Min
- Variance

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

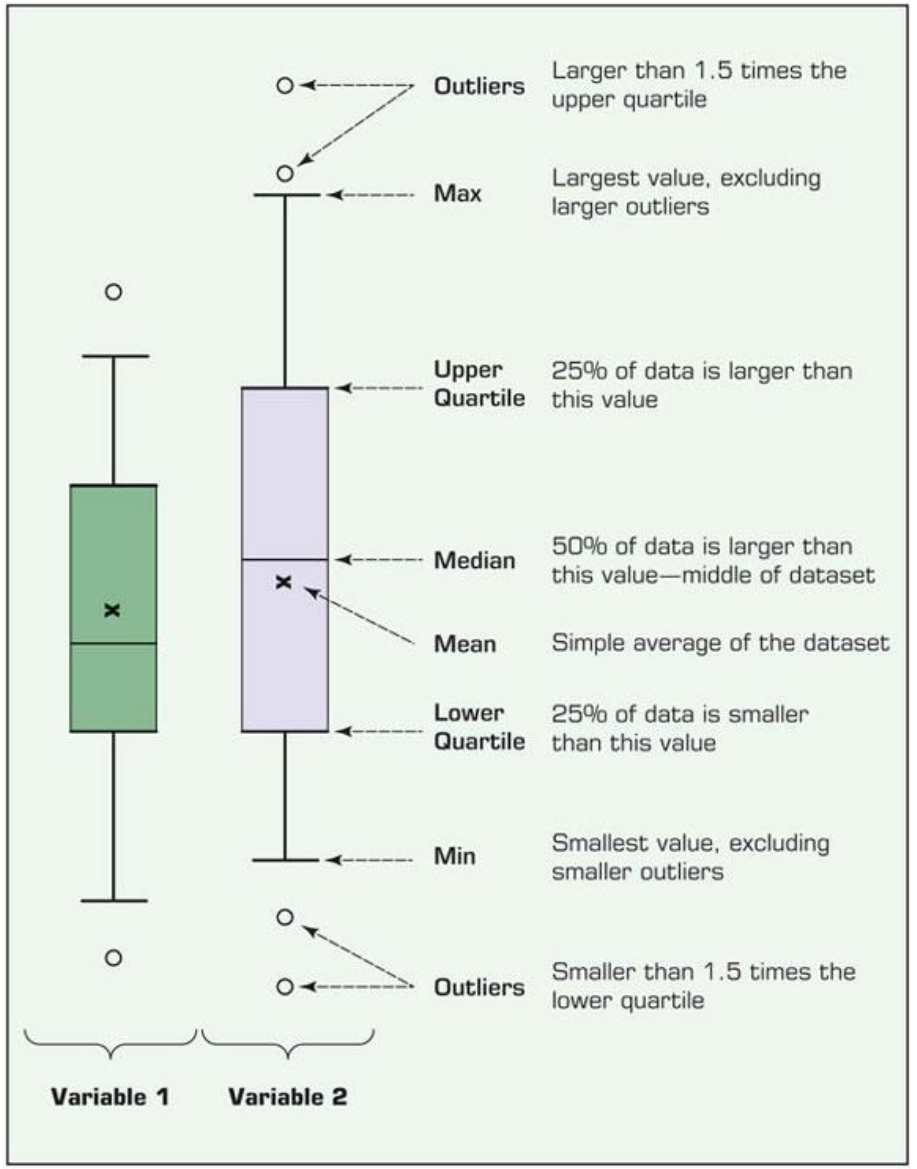
Standard Deviation

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

- Mean Absolute Deviation (MAD)
  - Average absolute deviation from the mean

Descriptive Statistics Measures of Dispersion (2 of 2)

- Box-and-Whiskers Plot
  - a.k.a. box-plot
- Quartiles:
  - Upper Q3
  - Middle Q2 = median
  - Lower Q1
- IQR Q3 - Q1
- Outliers borders
  - Lower = Q1- (1.5*IQR)
  - Upper = Q3 + (1.5*IQR)



Descriptive Statistics Measures of Centrality Tendency (2 of 2)

- Histogram - frequency chart
- Skewness
  - Measure of asymmetry

$$skewness = s = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)s^3}$$

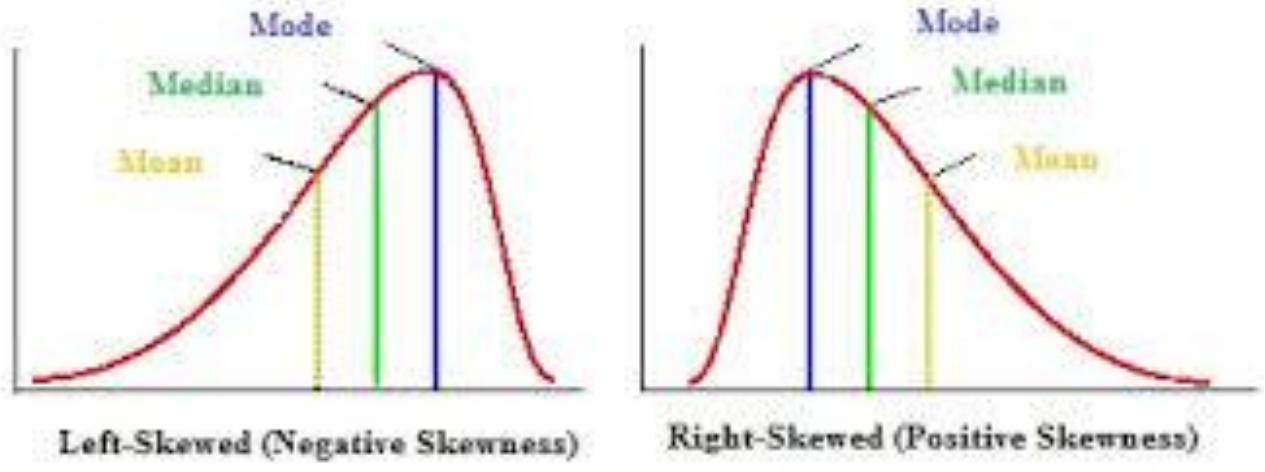
- Kurtosis
  - Peak/tall/skinny nature of the distribution

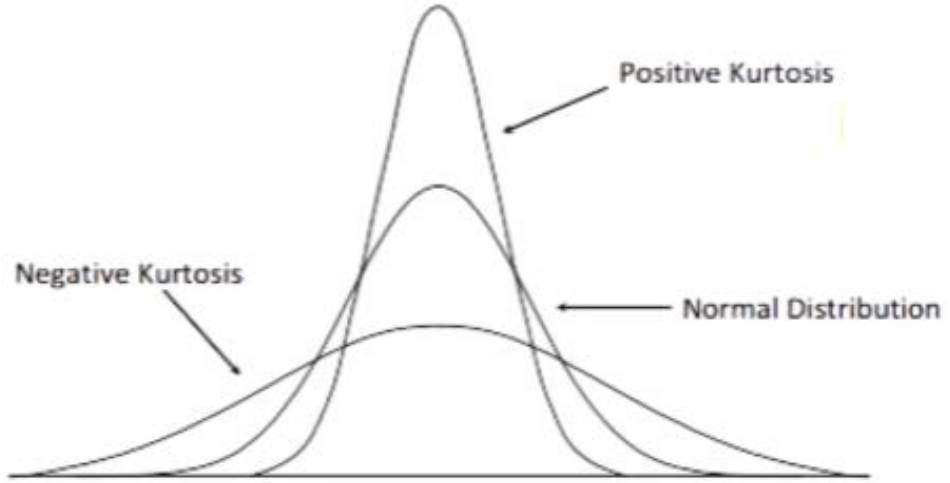
$$kurtosis = K = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} - 3$$

Skewness vs Kurtosis

Skewness

- =0 for Normal
- <0 skewed left
- >0 skewed right

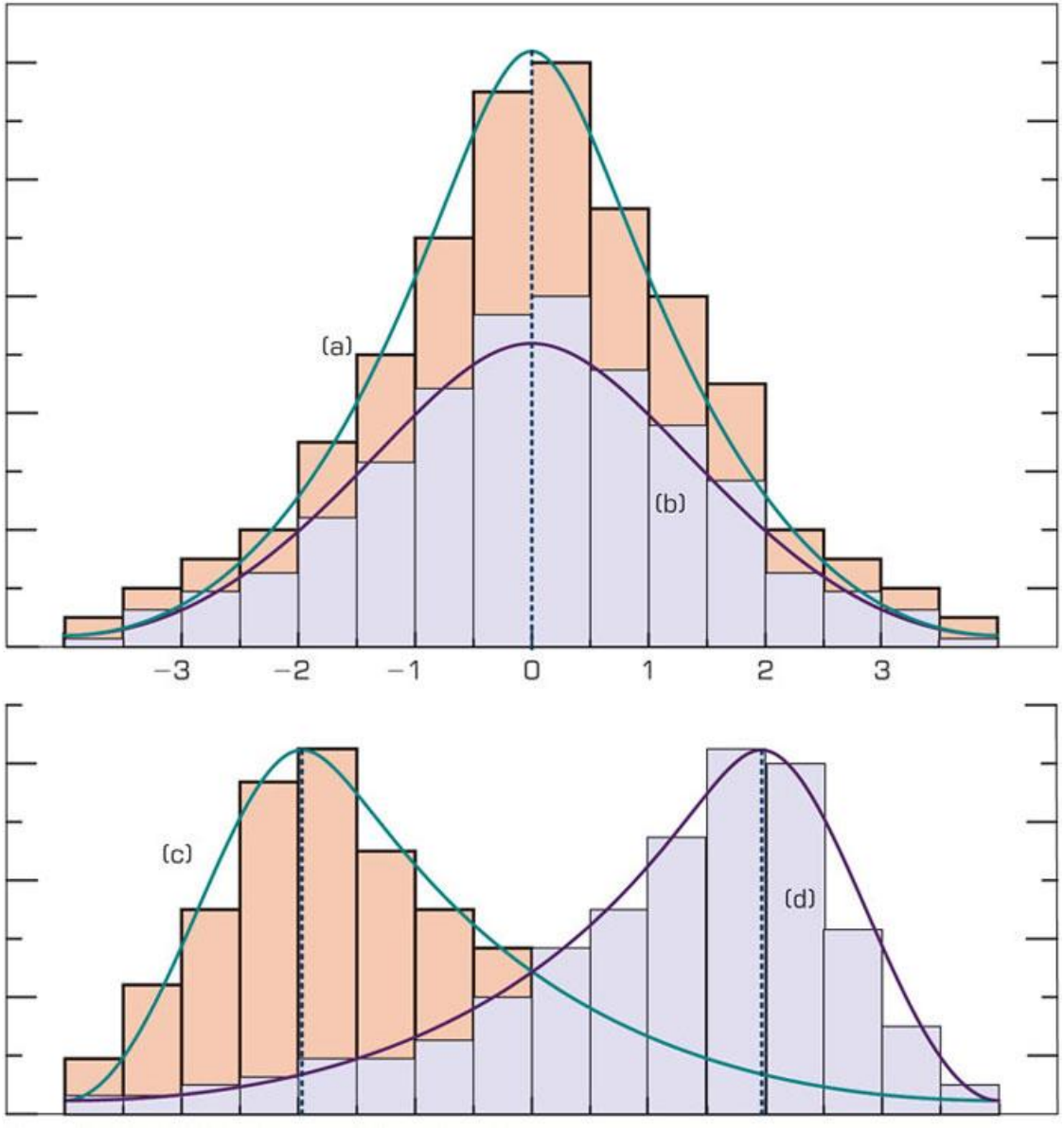




Kurtosis

- =3 for Normal
- >3 positive (thin)
- <3 negative (thick)

Relationship Between Dispersion and Shape Properties



Regression Modeling for Inferential Statistics

- **Regression**
  - A part of inferential statistics
  - The most widely known and used analytics technique in statistics
  - Used to characterize relationship between explanatory (input) and response (output) variable

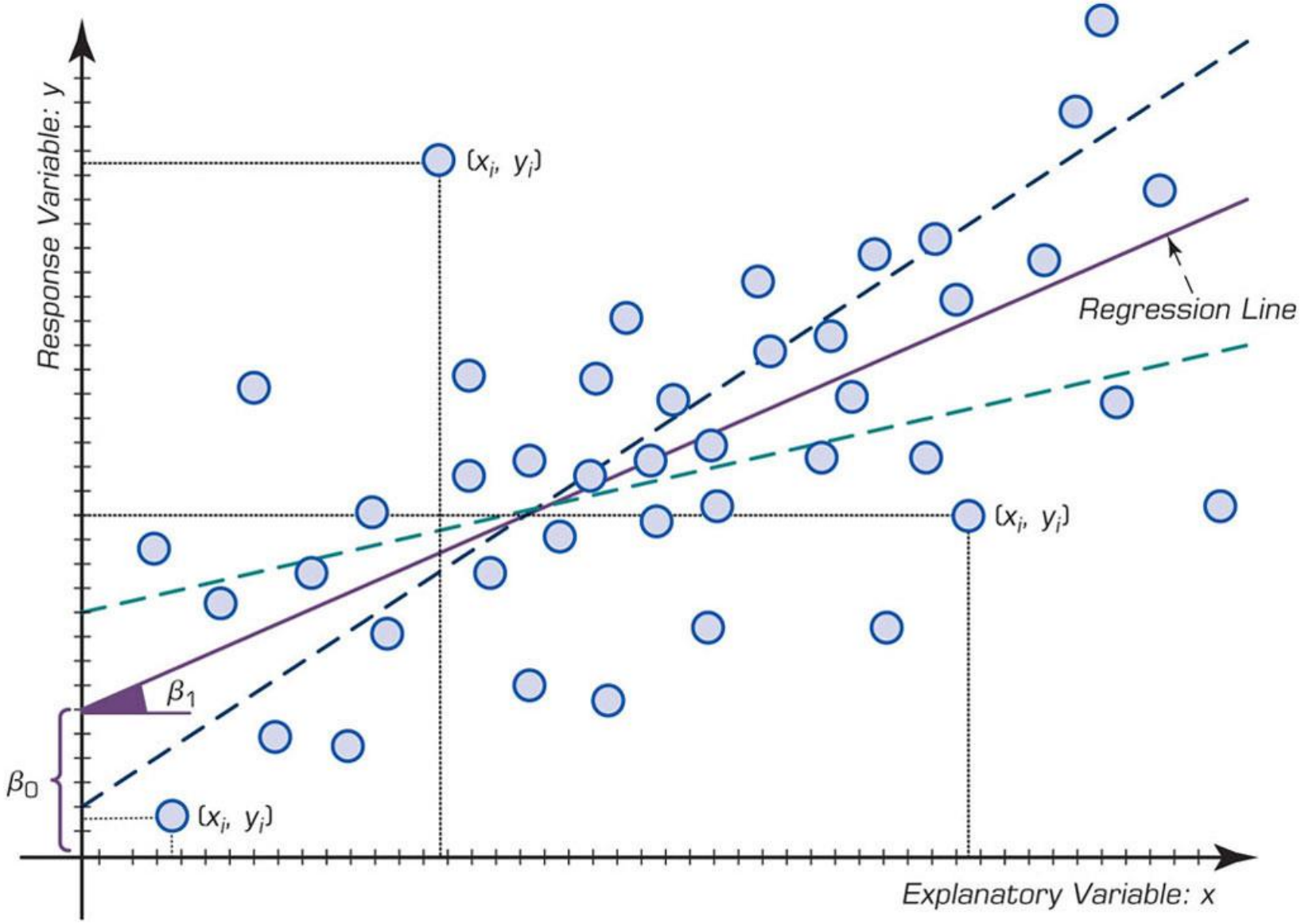
- **It can be used for**
  - Hypothesis testing (explanation)
  - Forecasting (prediction)

Regression Modeling (1 of 3)

- **Simple Regression versus Multiple Regression**
  - Base on number of input variables

- **How do we develop linear regression models?**
  - Scatter plots (visualization—for simple regression)
  - Ordinary least squares (OLS) method
    - A line that minimizes the sum of squared residuals.

Regression Modeling (2 of 3)



Regression Modeling (3 of 3)

- *x*: input, *y*: output
- Simple Linear Regression

$$y = \beta_0 + \beta_1 x$$

- Multiple Linear Regression

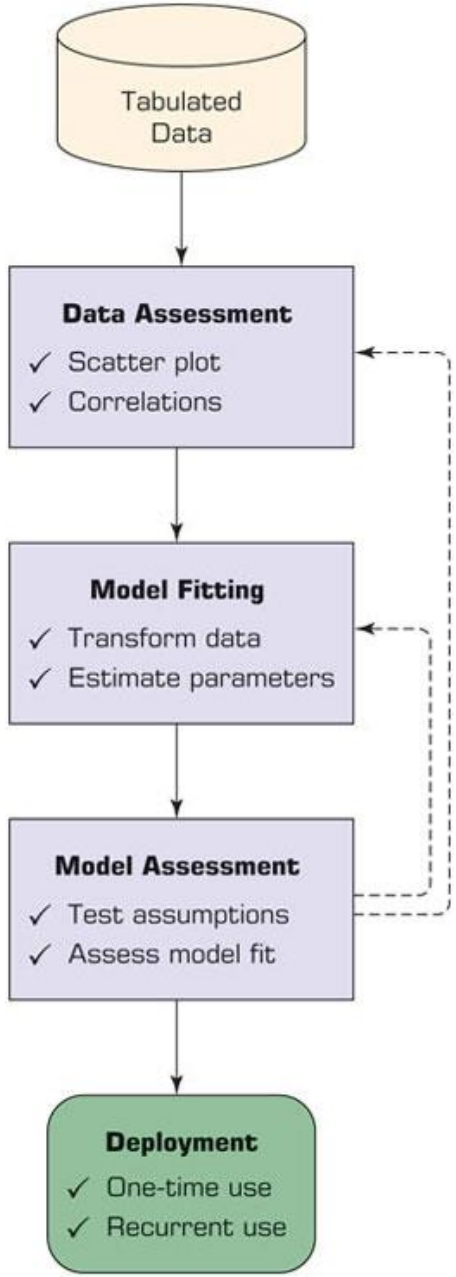
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$$

- The meaning of Beta (β) coefficients
  - Sign (+ or −) and magnitude

Process of Developing a Regression Model

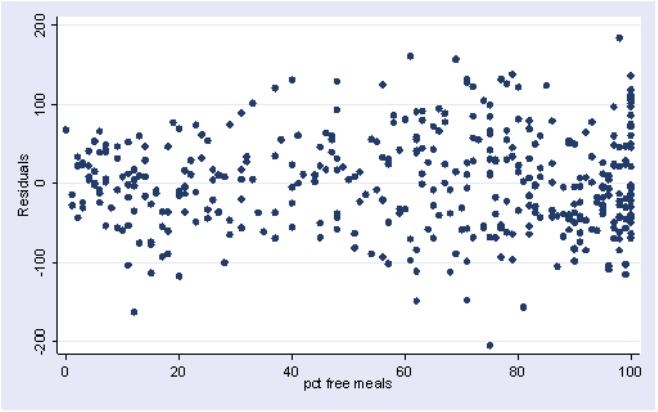
How do we know if the model is good enough?

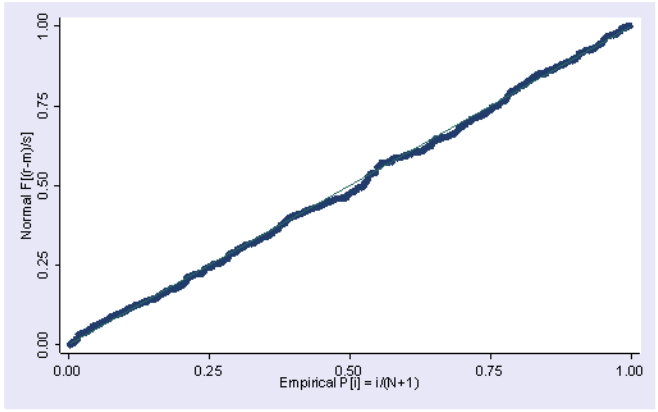
- R^2 (R-Square)
- *p* Values
- Error measures (for prediction problems)
  - MSE, MAD, RMSE

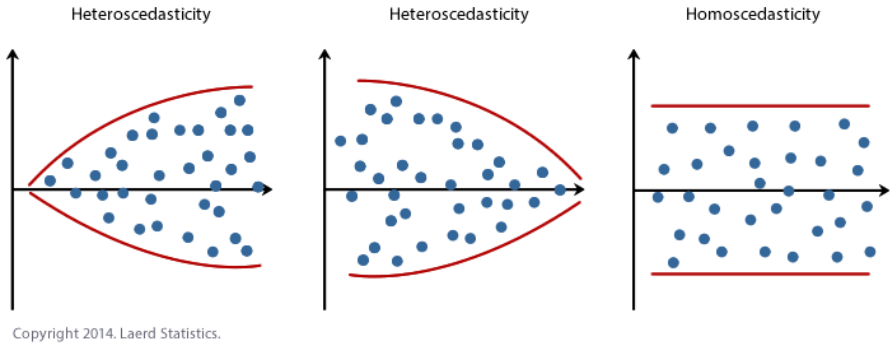


Regression Modeling Assumptions

- **Linearity** → linear relationship between x's & y.
  - Scatter plot (for simple regression)
  - Standardized residuals against each independent variable.
- **Independence** (of errors) → y errors are uncorrelated with each other.
- **Normality** (Normal Distribution) → errors are normally distrusted.
  - Kernel density plot.
  - Standardized normal probability (P-P)
- **Constant Variance or errors** (homoscedasticity) → No particular pattern of errors.
  - Residuals versus predicted values plot.
  - Breusch-Pagan test
- **Multicollinearity** → x's are not correlated.
  - Correlation.
  - VIF





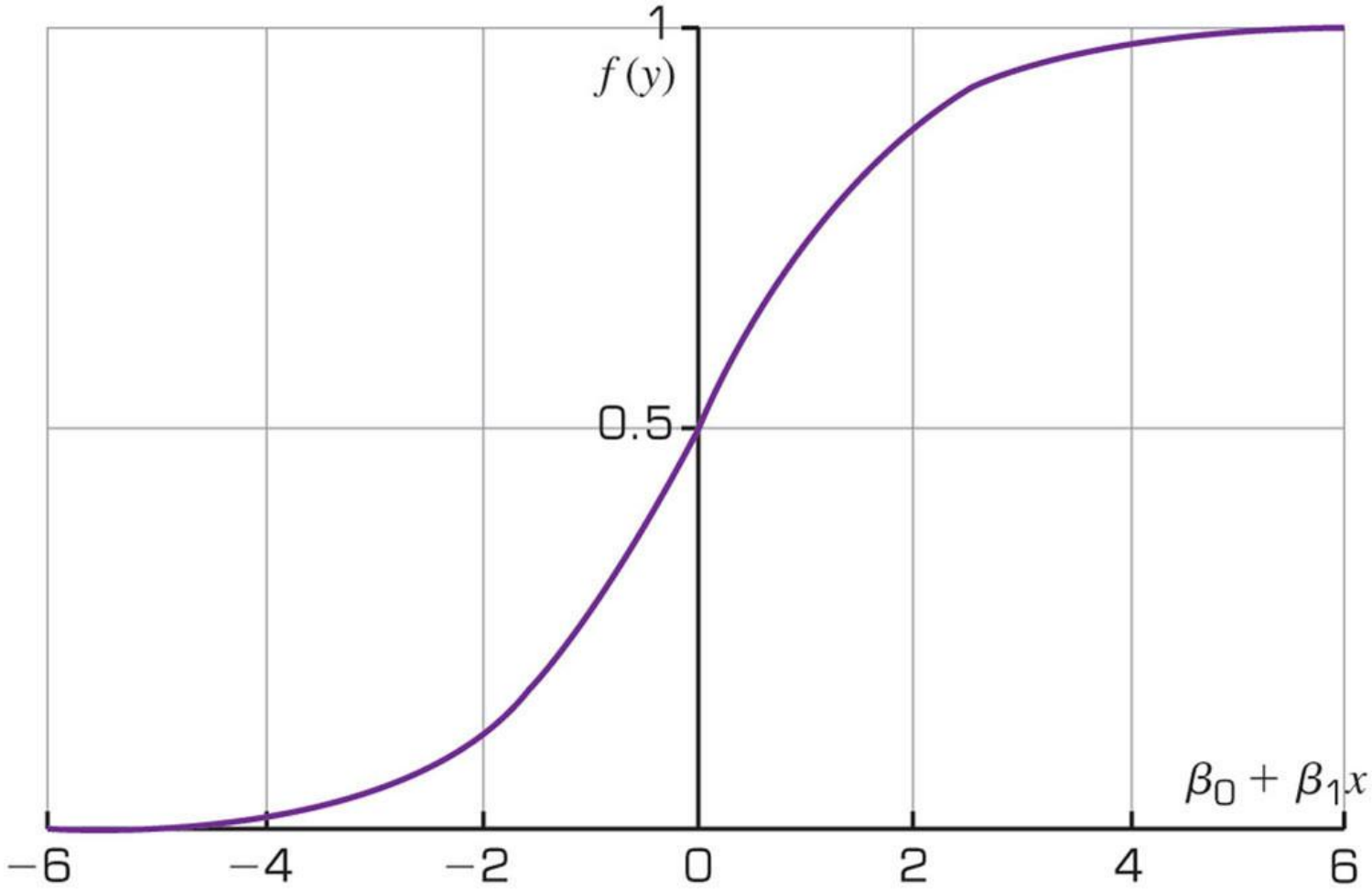


Logistic Regression Modeling (1 of 2)

- A very popular statistics-based classification algorithm
- Employs supervised learning
- Developed in 1940s
- The difference between Linear Regression and Logistic Regression
  - In Logistic Regression Output/Target variable is a binomial (binary classification) variable (as supposed to numeric variable)

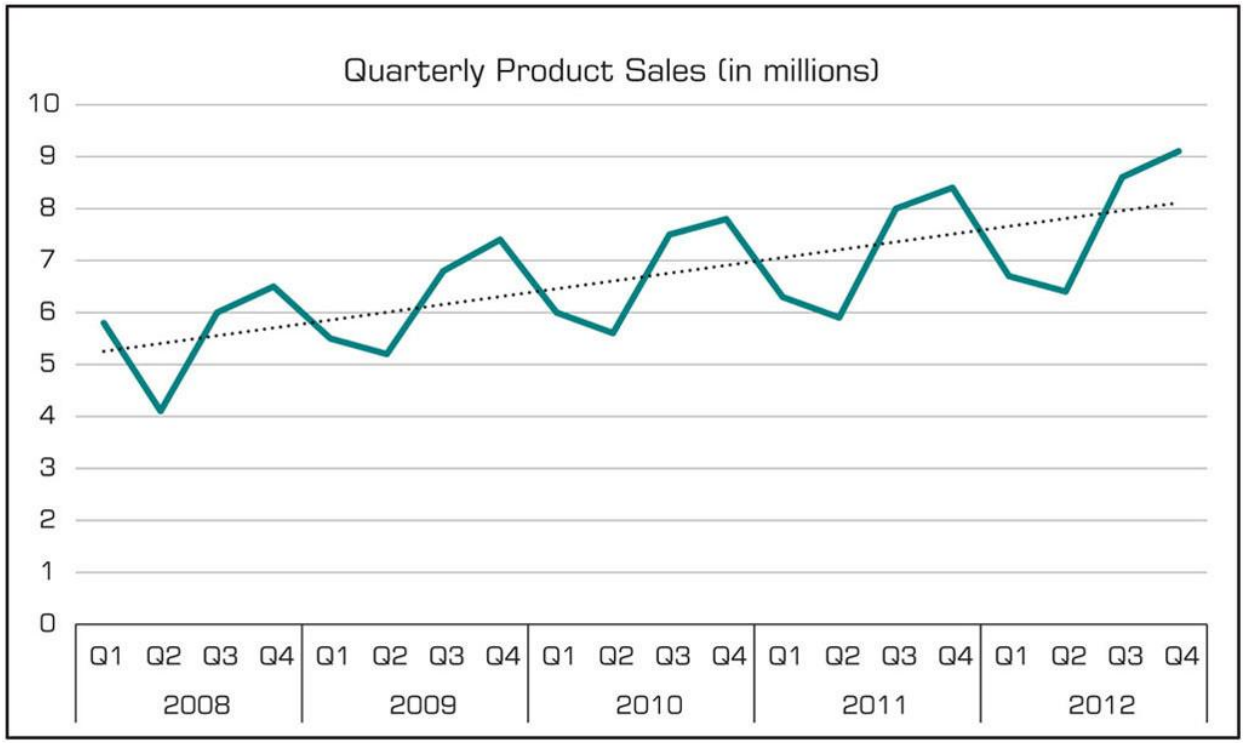
Logistic Regression Modeling (2 of 2)

$$f(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



Time Series Forecasting

- **Definition**: the use of mathematical modeling to predict future values of the variable of interest based on previously observed values.



Business Reporting Definitions and Concepts

- Report = Information → Decision
- Report?
  - Any communication artifact prepared to convey specific information

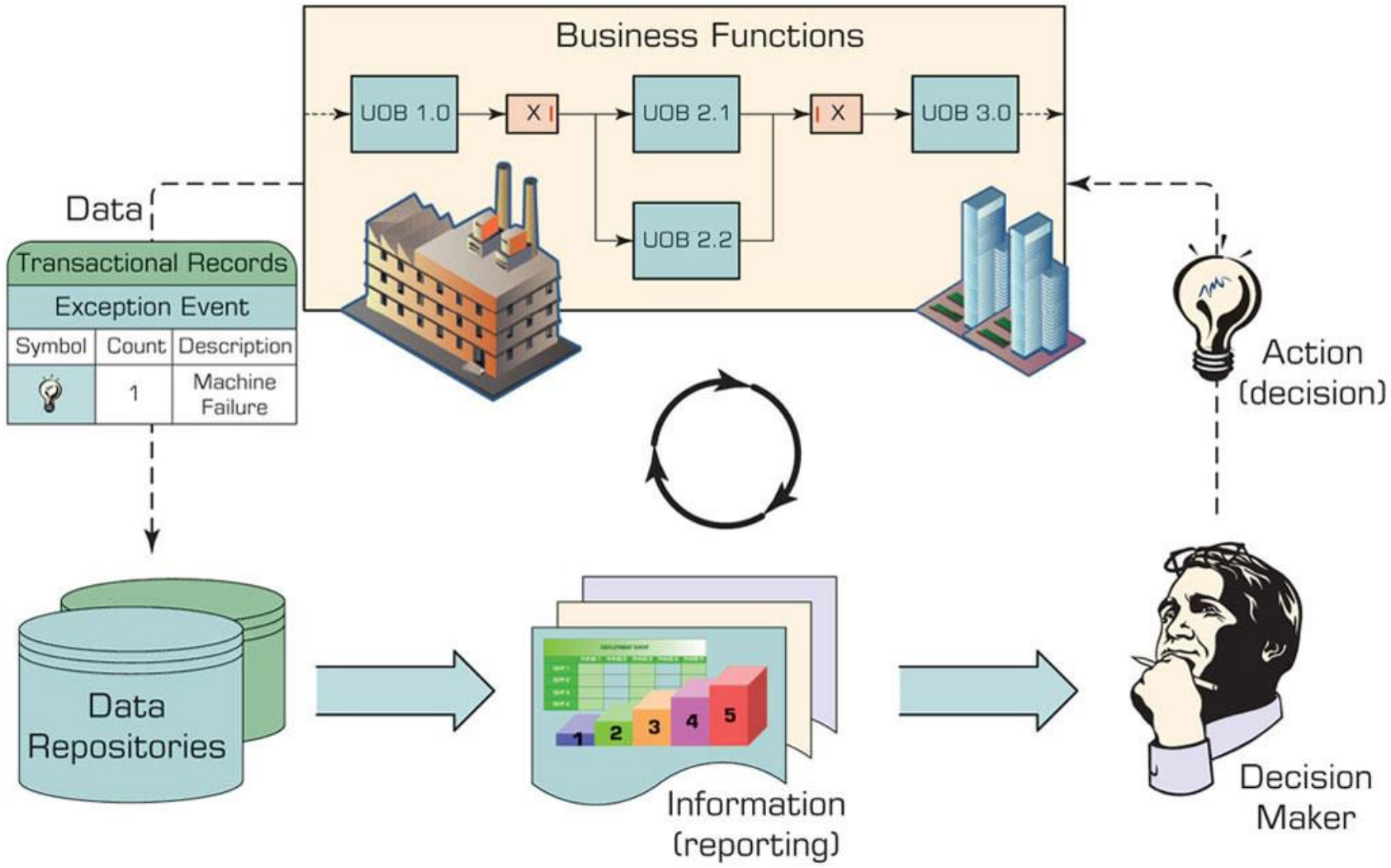
- A report can fulfill many functions
  - To ensure proper departmental functioning
  - To provide information
  - To provide the results of an analysis
  - To persuade others to act
  - To create an organizational memory…

What is a Business Report?

- A written document that contains information regarding business matters.
- Purpose: to improve managerial decisions
- Source: data from inside and outside the organization (via the use of ETL)
- Format: text + tables + graphs/charts
- Distribution: in-print, email, portal/intranet

Data acquisition → Information generation → Decision making → Process management

Business Reporting

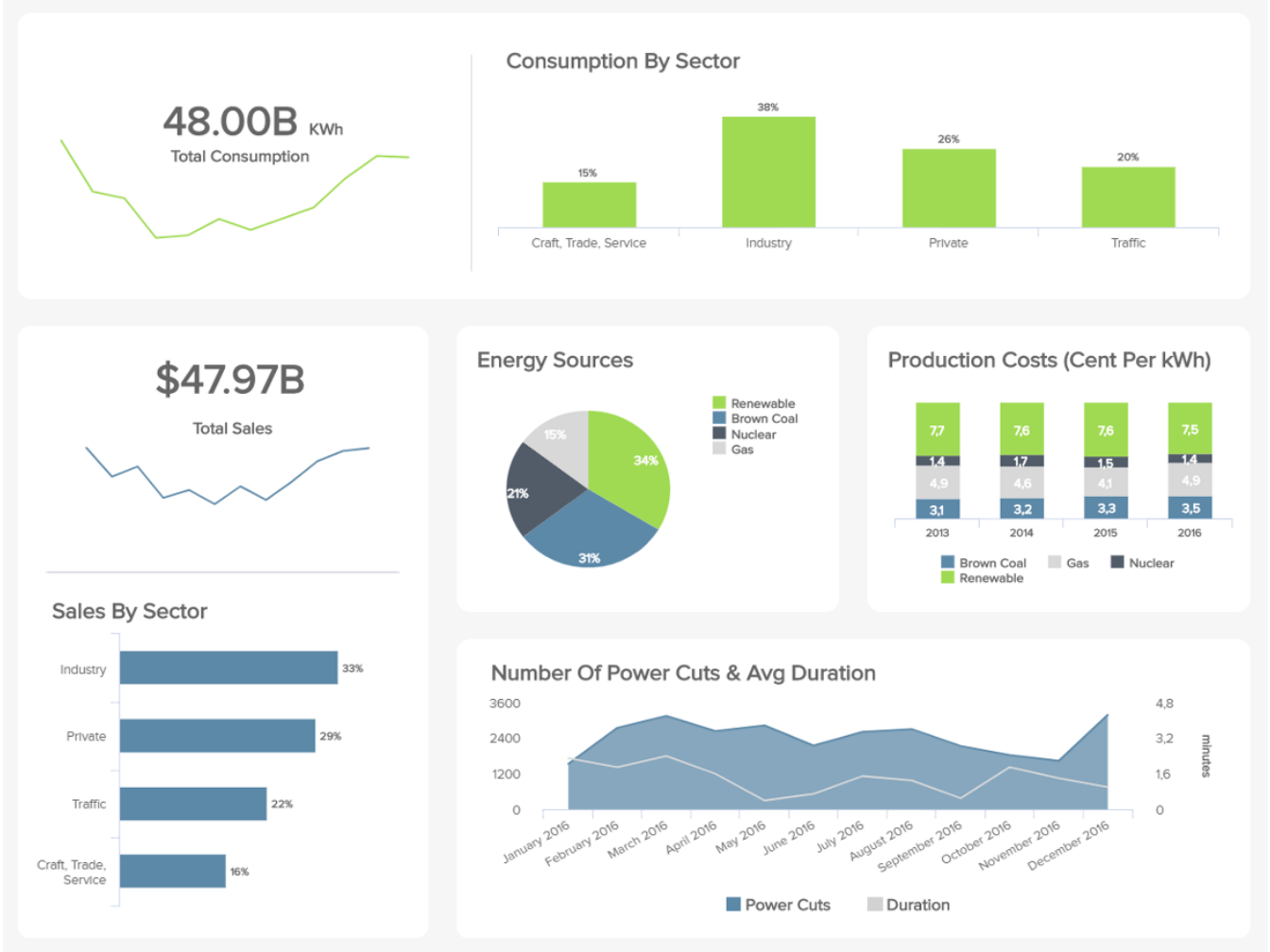


Types of Business Reports

- **Metric Management Reports**
  - Help manage business performance through metrics (SLAs for externals; K PIs for internals)
  - Can be used as part of Six Sigma and/or TQM

- **Dashboard-Type Reports**
  - Graphical presentation of several performance indicators in a single page using dials/gauges

- **Balanced Scorecard–Type Reports**
  - Include financial, customer, business process, and learning & growth indicators



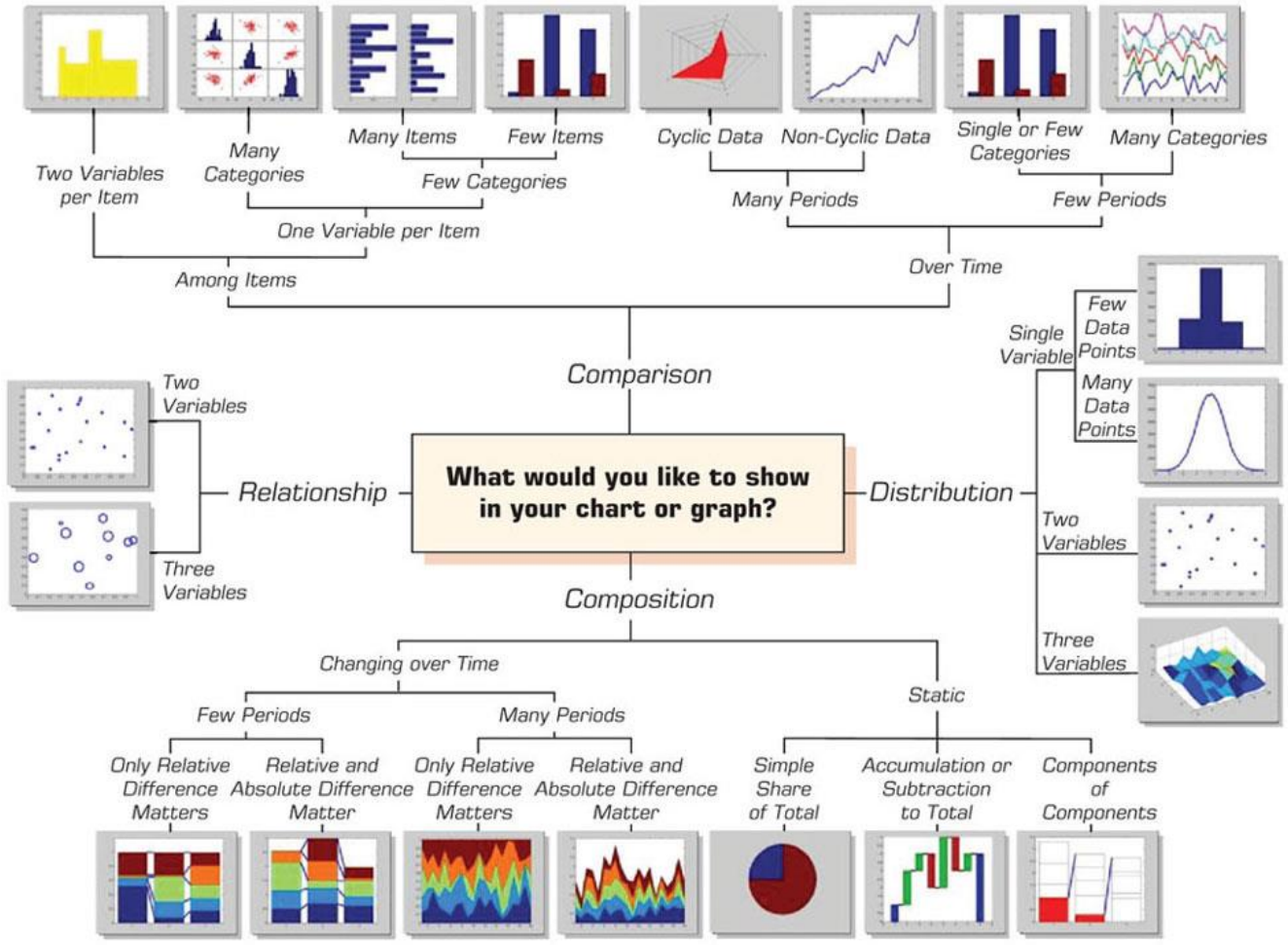
Data Visualization

Definition: “The use of visual representations to explore, make sense of, and communicate data.”

- Data visualization vs. Information visualization
- Information = aggregation, summarization, and contextualization of data
- Related to information graphics, scientific visualization, and statistical graphics
- Often includes charts, graphs, illustrations, …

Which Chart or Graph Should You Use?

Figure 3.21 A Taxonomy of Charts and Graphs.



Source: Adapted from Abela, A. (2008). *Advanced Presentations by Design: Creating Communication That Drives Action*. New York: Wiley.

Visual Analytics

- **A recently coined term**
  - Information visualization + predictive analytics
  - Embedding analytics capabilities into high-performance data visualization environment

- **Information visualization**
  - Descriptive, backward focused
  - “what happened” “what is happening”

- **Predictive analytics**
  - Predictive, future focused
  - “what will happen” “why will it happen”

- There is a strong move toward visual analytics

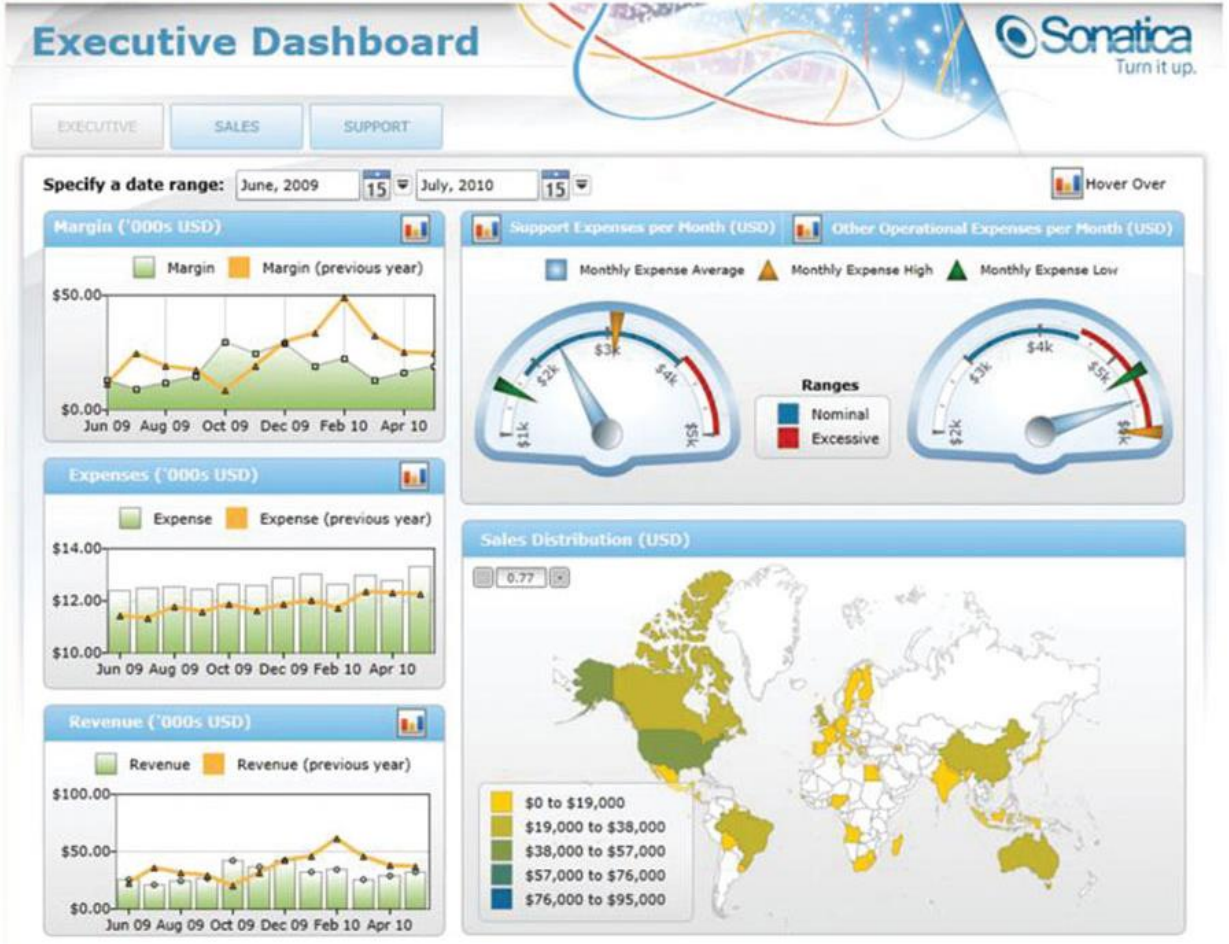
Performance Dashboards (1 of 4)

- Performance dashboards are commonly used in BPM software suites and BI platforms

- **Dashboards:**
  - Provide visual displays of important information.
  - Consolidated and arranged on a single screen.
  - So that information can be digested at a single glance and easily drilled in and further explored.

Performance Dashboards (2 of 4)

Figure 3.27 A Sample Executive Dashboard.



Source: A Sample Executive Dashboard from Dundas Data Visualization, Inc., **www.dundas.com,** reprinted with permission.

Performance Dashboards (3 of 4)

- **Dashboard design:**
  - The fundamental challenge of dashboard design is to display all the required information on a single screen, clearly and without distraction, in a manner that can be assimilated quickly

- **Three layers of information**
  - *Monitoring* → graphical, abstracted data to monitor KPIs.
  - *Analysis* → summarized dimensional data to analyze the root cause of problems.
  - *Management* →detailed operational data identifying what actions to take.

Performance Dashboards (4 of 4)

- **What to look for in a dashboard:**
  - Use of visual components to highlight data and exceptions that require action.
  - Transparent to the user, meaning that they require minimal training and are extremely easy to use
  - Combine data from a variety of systems into a single, summarized, unified view of the business
  - Enable drill-down or drill-through to underlying data sources or reports
  - Present a dynamic, real-world view with timely data
  - Require little coding to implement, deploy, and maintain

Best Practices in Dashboard Design

- Benchmark KPIs with Industry Standards: best practice KPIs
- Wrap the Metrics with Contextual Metadata: characteristics of data
- Validate the Design by a Usability Specialist: user friendly
- Prioritize and Rank Alerts and Exceptions: RGB
- Enrich Dashboard with Business-User Comments: add comments
- Present Information in Three Different Levels: drill-down
- Pick the Right Visual Constructs: show the desired plot
- Provide for Guided Analytics: help

Q & A