

# What is Data Mining?

- ▶ “...the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data...” (Gartner Group)
- ▶ “...the analysis of observational data sets to find unsuspected relationships and to summarize data in novel ways...” (Hand et al.)
- ▶ “...is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization...” (Cabana et al.)

# The Need for Human Direction of Data Mining

- ▶ Some early data mining definitions described process as “automatic”
- ▶ “...this has misled many people into believing data mining is product that can be bought rather than a discipline that must be mastered.” (Berry, Linoff)
- ▶ Automation no substitute for human input
- ▶ Data mining is easy to do badly
- ▶ Understanding statistical and mathematical model structures of underlying software required
- ▶ Humans need to be actively involved in every phase of data mining process
- ▶ Task of data mining should be integrated into human process of problem solving

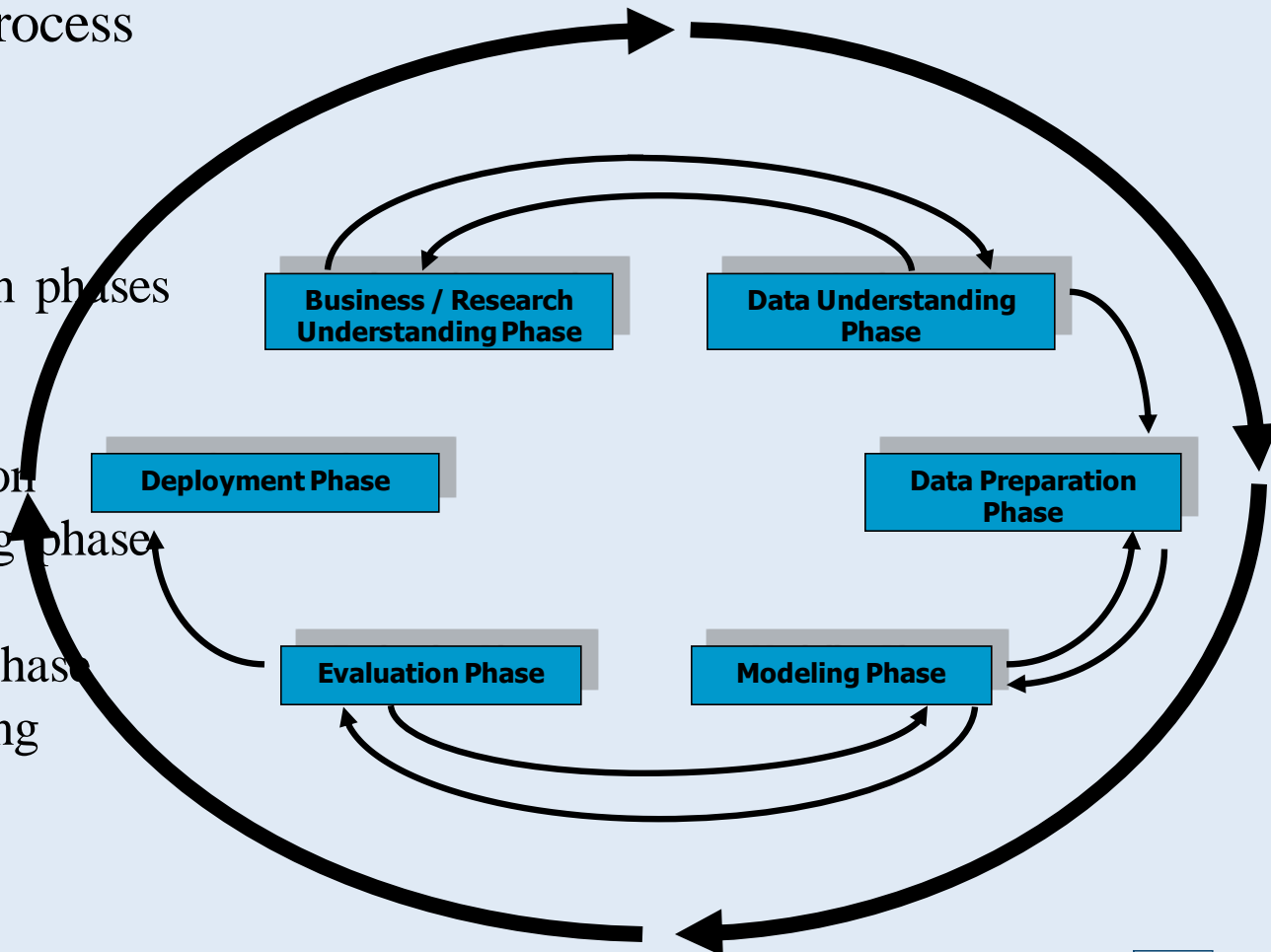
# Cross Industry Standard Process: CRISP-DM

- ▶ Cross-Industry Standard Process for Data Mining (CRISP-DM) developed in 1996
- ▶ Contributors include DaimlerChrysler, SPSS, and NCR
- ▶ Developed to fit data mining into general business strategy
- ▶ Process vendor and tool-neutral
- ▶ Non-proprietary and freely available
- ▶ Data mining projects follow iterative, adaptive life cycle consisting of 6 phases
- ▶ Phase sequences are adaptive
- ▶ Next, Figure 1.1 illustrates CRISP-DM lifecycle

# Cross Industry Standard Process: CRISP-DM (cont'd)

■ Iterative CRISP-DM process shown in outer circle

- ▶ Most significant dependencies between phases shown
- ▶ Next phase depends on results from preceding phase
- ▶ Returning to earlier phase possible before moving forward



# Cross Industry Standard Process: CRISP-DM (cont'd)

## ■ (1) Business/Research Understanding Phase

- ▶ Define business/research requirements and objectives
- ▶ Translate objectives into data mining problem definition
- ▶ Prepare initial strategy to meet objectives

## ■ (2) Data Understanding Phase

- ▶ Collect the data
- ▶ Assess data quality
- ▶ Perform exploratory data analysis (EDA)

# Cross Industry Standard Process: CRISP-DM (cont'd)

## ■ (3) Data Preparation Phase

- ▶ Cleanse, prepare, and transform data set
- ▶ Prepares for modeling in subsequent phases
- ▶ Select cases and variables appropriate for analysis

# Cross Industry Standard Process: CRISP-DM (cont'd)

## ■ (4) Modeling Phase

- ▶ Select and apply one or more modeling techniques
- ▶ Calibrate model settings to optimize results
- ▶ If necessary, additional data preparation may be required

## ■ (5) Evaluation Phase

- ▶ Evaluate one or more models for effectiveness
- ▶ Determine whether defined objectives achieved
- ▶ Make decision regarding data mining results before deploying to field

# Cross Industry Standard Process: CRISP-DM (cont'd)

## ■ (6) Deployment Phase

- ▶ Make use of models created
- ▶ Simple deployment: generate report
- ▶ Complex deployment: implement additional data mining effort in another department
- ▶ In business, customer often carries out deployment based on model

See <http://www.crisp-dm.org> for more information



# What Tasks Can Data Mining Accomplish?

- ▶ For example, those laid-off now less financially secure; therefore, prefer alternate candidate
- ▶ Data mining models should be transparent
- ▶ That is, results should be interpretable by humans
- ▶ Some data mining methods more transparent than others
- ▶ For example, Decision Trees (transparent) <-> Neural Networks (opaque)
- ▶ High-quality description accomplished using Exploratory Data Analysis (EDA)
- ▶ Graphical method of exploring patterns and trends in data

# What Tasks Can Data Mining Accomplish?

## ■ Six common data mining tasks

- ▶ Description
- ▶ Estimation
- ▶ Prediction
- ▶ Classification
- ▶ Clustering
- ▶ Association

## ■ (1) Description

- ▶ Describes patterns or trends in data
- ▶ For example, pollster may uncover patterns suggesting those laid-off less likely to support incumbent
- ▶ Descriptions of patterns, often suggest possible explanations

# What Tasks Can Data Mining Accomplish? (cont'd)

## ■ (2) Estimation

- ▶ Similar to Classification task, except target variable numeric
- ▶ Models built from complete data records
- ▶ Records include values for each predictor field and numeric target variable in training set
- ▶ For new observations, estimate of target variable made
  
- ▶ For example, estimate a patient's systolic blood pressure, based on patient's age, gender, body-mass index, and sodium levels

Here, estimation model built from training set records

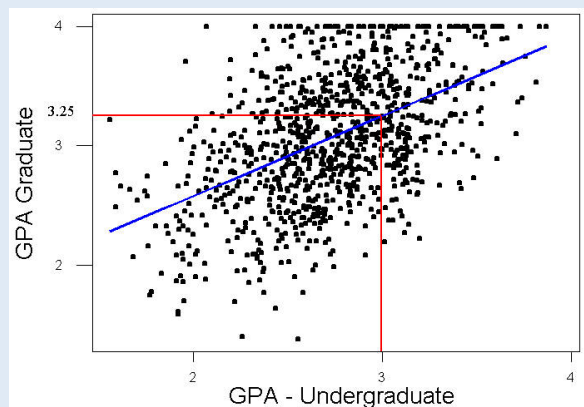
Model then estimates value for new case

# What Tasks Can Data Mining Accomplish? (cont'd)

- ▶ **Estimation Tasks in Business and Research:**
- ▶ Estimate amount of money, family of four will spend on back-to-school shopping
- ▶ Estimate percentage decrease in rotary movement sustained to NFL player with knee injury
- ▶ Estimate number of points basketball player scores when double-teamed in playoffs
- ▶ Estimate GPA of graduate student, based on student's undergraduate GPA

# What Tasks Can Data Mining Accomplish? (cont'd)

- ▶ Figure 1.2 shows scatter plot of graduate GPA against undergraduate GPA



- ▶ Linear regression finds line (blue) best approximating relationship between two variables
- ▶ Regression line estimates student's graduate GPA based on their undergraduate GPA

# What Tasks Can Data Mining Accomplish? (cont'd)

- ▶ Minitab statistical software produces regression equation  $\hat{y} = 1.24 + 0.67x$
- ▶ Therefore, estimated student's graduate GPA = 1.24 plus 0.67 times their undergraduate GPA
- ▶ For example, suppose student's undergraduate GPA = 3.0
- ▶ According to estimation model
- ▶ Estimated student's graduate GPA =  $1.24 + 0.67(3.0) = 3.25$
- ▶ Point  $(x = 3.0, \hat{y} = 3.25)$  lies on regression line
- ▶ Statistical Analysis uses several estimation methods: point estimation, confidence interval estimation, linear regression and correlation, and multiple regression

# What Tasks Can Data Mining Accomplish? (cont'd)

## ■ (3) Prediction

- ▶ Similar to classification and estimation, except results lie in the future
- ▶ Predict price of stock 3 months into future, based on past performance



# What Tasks Can Data Mining Accomplish? (cont'd)

- ▶ Predict percentage increase in traffic deaths next year, if speed limit increased
- ▶ Predict whether molecule in newly discovered drug leads to profitable pharmaceutical drug

Methods used for classification and prediction applicable to prediction  
Includes point estimation, confidence interval estimation, linear regression and correlation, and multiple regression



# What Tasks Can Data Mining Accomplish? (cont'd)

## ■ (4) Classification

- ▶ Classification requires categorical target variable such as Income Bracket
- ▶ Three values include “High”, “Middle”, “Low”
- ▶ Data model examines records containing input fields and target field
- ▶ Table shows several records from data set

Subject	Age	Gender		Income Bracket
001	47	F	Software Engineer	High
002	28	M	Marketing Consultant	Middle
003	35	M	Unemployed	Low
...	...	...	...	...

# What Tasks Can Data Mining Accomplish? (cont'd)

- ▶ Records of persons in data set used to “train” classification model
- ▶ First, Model built from data records, where value of categorical target variable (Income Bracket) already known
- ▶ Algorithm “first learns about” which combinations of input fields are associated with Income Bracket values in training set
- ▶ For example, algorithm may determine that older females associated with high income
  
- ▶ Next, trained model examines new records
- ▶ Information regarding Income Bracket not available

# What Tasks Can Data Mining Accomplish? (cont'd)

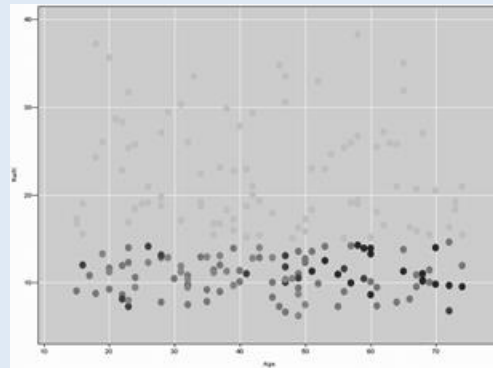
- ▶ Based on classifications in training set, new records classified
- ▶ For example, 63-year old female professor might be classified in “High” income bracket

## ■ Classification Tasks in Business and Research:

- ▶ Determine whether credit card transaction fraudulent
- ▶ Assessing mortgage application to determine “good” or “bad” credit risk
- ▶ Diagnosing whether particular disease present

# What Tasks Can Data Mining Accomplish? (cont'd)

- ▶ Determine if will written by deceased, or fraudulently by someone else
- ▶ Identify whether certain financial behavior represents terrorist threat



- ▶ Scatter plot shows Na/K ratio against Age for 200 patients
- ▶ For example, classify drug type to prescribe based on patient's age and sodium/potassium ratio

# What Tasks Can Data Mining Accomplish? (cont'd)

- ▶ Actual drug type prescribed symbolized by shade (light, medium, dark) of points
- ▶ Suppose prescription of new patient based on this data set?
- **Prescribe which drug for young patient with high Na/K ratio?**
  - ▶ Young patients plotted on left
  - ▶ High Na/K plotted on upper-half
  - ▶ Quadrant of graph shows light points
  - ▶ Recommended drug = Y (corresponds to light points)
- **Prescribe which drug for older patient with low Na/K ratio?**
  - ▶ Lower-right half of graph shows patients prescribed different drug types

# What Tasks Can Data Mining Accomplish? (cont'd)

## ■ (5) Clustering

- ▶ Refers to grouping records into classes of similar objects
  - ▶ Clustering algorithm seeks to segment data set into homogeneous subgroups
  - ▶ Where similarity of records in clusters maximized, and similarity to records outside clusters minimized
  - ▶ Target variable not specified
- 
- ▶ For example, Claritas, Inc. PRIZM software clusters demographic profiles for different geographic areas according to zip code

# What Tasks Can Data Mining Accomplish? (cont'd)

- ▶ **Table shows 62 distinct “lifestyle” types used by PRIZM**

01 Blue Blood Estates	02 Winner's Circle	03 Executive Suites	04 Pools & Patios
05 Kids & Cul-de-Sacs	06 Urban Gold Coast	07 Money & Brains	08 Young Literati
09 American Dreams	10 Bohemian Mix	11 Second City Elite	12 Upward Bound
13 Gray Power	14 Country Squires	15 God's Country	16 Big Fish, Small Pond
17 Greenbelt Families	18 Young Influentials	19 New Empty Nests	20 Boomers & Babies
21 Suburban Sprawl	22 Blue-Chip Blues	23 Upstarts & Seniors	24 New Beginnings
25 Mobility Blues	26 Gray Collars	27 Urban Achievers	28 Big City Blend
29 Old Yankee Rows	30 Mid-City Mix	31 Latino America	32 Middleburg Managers
33 Boomtown Singles	34 Starter Families	35 Sunset City Blues	36 Towns & Gowns
37 New Homesteaders	38 Middle America	39 Red, White & Blues	40 Military Quarters
41 Big Sky Families	42 New Eco - topia	43 River City, USA	44 Shotguns & Pickups
45 Single City Blues	46 Hispanic Mix	47 Inner Cities	48 Smalltown Downtown
49 Hometown Retired	50 Family Scramble	51 Southside City	52 Golden Ponds
53 Rural Industria	54 Norma Rae-Ville	55 Mines & Mills	56 Agri - Business
57 Grain Belt	58 Blue Highways	59 Rustic Elders	60 Back Country Folks
61 Scrub Pine Flats	62 Hard Scrabble		

# What Tasks Can Data Mining Accomplish? (cont'd)

- ▶ What do the clusters mean?
- ▶ According to PRIZM, Clusters for Beverly Hills, CA 90210 include:
  - Cluster 01: Blue Blood Estates
  - Cluster 10: Bohemian Mix
  - Cluster 02: Winner's Circle
  - Cluster 08: Young Literati
- ▶ Description of Cluster 01, "...old money' heirs that live in America's wealthiest suburbs...accustomed to privilege and live luxuriously..."



# What Tasks Can Data Mining Accomplish? (cont'd)

## ■ Clustering Tasks in Business and Research:

- ▶ Target marketing niche product for small business that does not have large marketing budget
- ▶ For accounting purposes, segment financial behavior into benign and suspicious categories
- ▶ Use as dimensionality-reduction tool for data set having several hundred inputs
- ▶ Determine gene expression clusters, where many genes exhibit similar behavior or characteristics
- ▶ Clustering often used as preliminary step in data mining
- ▶ Resulting clusters used as input to different technique downstream, such as neural networks

# What Tasks Can Data Mining Accomplish? (cont'd)

## ■ (6) Association

- ▶ Find out which attributes “go together”
- ▶ Market Basket Analysis commonly used in business applications
- ▶ Quantify relationships in the form of Rules

## ■ IF antecedent THEN consequent

- ▶ Rules measured using support and confidence
- ▶ For example, discover which items in supermarket are purchased together
- ▶ Thursday night 200 of 1,000 customers bought diapers, and of those buying diapers, 50 purchased beer
- ▶ Association Rule: “IF buy diapers, THEN buy beer”
- ▶ Support =  $200/1,000 = 5\%$ , and confidence =  $50/200 = 25\%$

# What Tasks Can Data Mining Accomplish? (cont'd)

## ► Association Tasks in Business and Research:

- Investigating proportion of subscribers to cell phone plan responding positively to service upgrade offer
- Predicting degradation in telecommunication networks
- Discovering which items in supermarket purchased together
- Determining proportion of cases where administering new drug exhibits serious side effects

Two commonly-used algorithms for generating association rules  
A Priori and Generalized Rule Induction (GRI)

# Why Do We Preprocess Data?

## Data Understanding, Data Preparation

- ▶ Raw data often incomplete, noisy
- ▶ May contain:
  - Obsolete fields
  - Missing values
  - Outliers
  - Data in a form not suitable for data mining
  - Erroneous values

# Why Do We Preprocess Data? (cont'd)

- **For data mining purposes, database values must undergo data cleaning and data transformation**

- ▶ Data often from legacy databases where values:

- Not looked at in years
- Expired
- No longer relevant
- Missing
- Minimize GIGO (Garbage In Garbage Out)
- IF garbage input minimized → THEN garbage in results minimized
- Data preparation is 60% of effort for data mining process (Pyle)

# Data Cleaning

## Data errors in Table 2.1 examined:

### ■ Five-numeral U.S. Zip Code?

- ▶ Not all countries use same zip code format, 90210 (U.S.) vs. J2S7K7 (Canada)
- ▶ Should expect unusual values in some fields
- ▶ For example, global commerce

### ■ Four Digit Zip Code?

- ▶ Leading zero truncated, 6269 vs. 06269 (New England states)
- ▶ Database field numeric and chopped-off leading zero

# Data Cleaning (cont'd)

**TABLE 2.1** Can You Find Any Problems in This Tiny Data Set?

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	75000	C	M	5000
1002	J2S7K7	F	-40000	40	W	4000
1003	90210		10000000	45	S	7000
1004	6269	M	50000	0	S	1000
1005	55101	F	99999	30	D	3000

# Data Cleaning (cont'd)

## ■ Income Field Contains \$10,000,000?

- ▶ Assumed to measure gross annual income
- ▶ Possibly valid
- ▶ Still considered outlier (extreme data value)
- ▶ Some statistical and data mining methods affected by outliers

## ■ Income Field Contains -\$40,000?

- ▶ Income less than \$0?
- ▶ Value beyond bounds for expected income, therefore an error
- ▶ Caused by data entry error?
- ▶ Discuss anomaly with database administrator



# Data Cleaning (cont'd)

## ■ Income Field Contains \$99,999?

- ▶ Other values appear rounded to nearest \$5,000
- ▶ Value may be completely valid
- ▶ Value represents database code used to denote missing value?
- ▶ Confirm values in expected unit of measure, such as U.S. dollars
- ▶ Which unit of measure for income?
- ▶ Customer with zip code J2S7K7 in Canadian dollars?
- ▶ Discuss anomaly with database administrator

# Data Cleaning (cont'd)

## ▶ Age Field Contains "C"?

- ▶ Other records have numeric values for field
- ▶ Record categorized into group labeled "C"
- ▶ Value must be resolved
- ▶ Data mining software expects numeric values for field

## ■ Age Field Contains 0?

- ▶ Zero-value used to indicate missing/unknown value?
- ▶ Customer refused to provide their age?

# Data Cleaning (cont'd)

## ■ Age Field?

- ▶ Date-type fields may become obsolete
- ▶ Use date of birth, then derive Age

## ■ Marital Status Field Contains “S”?

- ▶ What does this symbol mean?
- ▶ Does “S” imply single or separated?
- ▶ Discuss anomaly with database administrator

# Handling Missing Data

- ▶ Missing values pose problems to data analysis methods
- ▶ More common in databases containing large number of fields
- ▶ Absence of information rarely beneficial to task of analysis
- ▶ In contrast, having more data almost always better
- ▶ Careful analysis required to handle issue

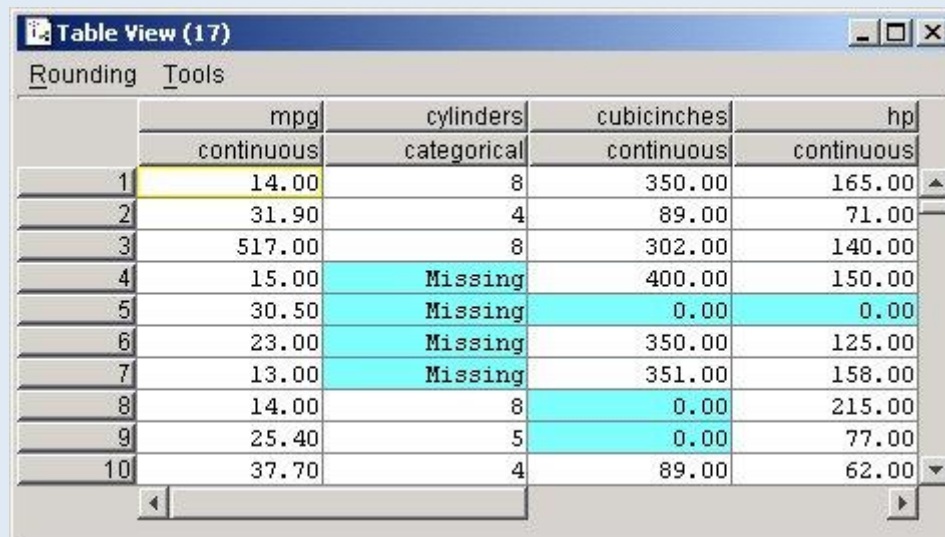
# Handling Missing Data (cont'd)

- **Examine *cars* dataset containing records for 261 automobiles manufactured in 1970s and 1980s**
- **Delete Records Containing Missing Values?**
  - ▶ Not necessarily best approach
  - ▶ Pattern of missing values may be systematic
  - ▶ Deleting records creates biased subset
  - ▶ Valuable information in other fields lost
- **Three Alternate Methods Available**
  - ▶ Insightful Miner (<http://www.insightful.com>) specifies method to replace values

# Handling Missing Data (cont'd)

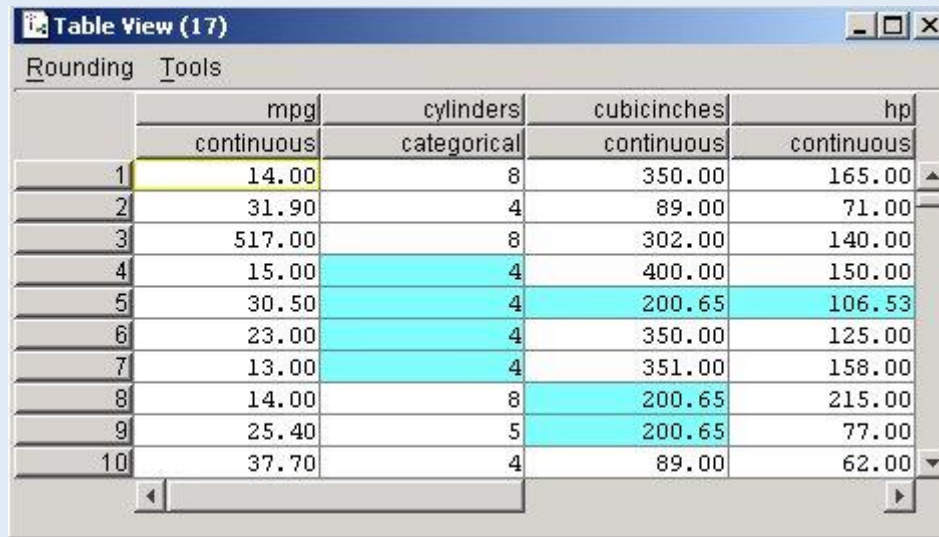
## ■ (1) Replace Missing Values with User-defined Constant

- ▶ Missing numeric values replaced with 0.0
- ▶ Missing categorical values replaced with “Missing”



	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.00	8	350.00	165.00
2	31.90	4	89.00	71.00
3	517.00	8	302.00	140.00
4	15.00	Missing	400.00	150.00
5	30.50	Missing	0.00	0.00
6	23.00	Missing	350.00	125.00
7	13.00	Missing	351.00	158.00
8	14.00	8	0.00	215.00
9	25.40	5	0.00	77.00
10	37.70	4	89.00	62.00

# Handling Missing Data (cont'd)



	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.00	8	350.00	165.00
2	31.90	4	89.00	71.00
3	517.00	8	302.00	140.00
4	15.00	4	400.00	150.00
5	30.50	4	200.65	106.53
6	23.00	4	350.00	125.00
7	13.00	4	351.00	158.00
8	14.00	8	200.65	215.00
9	25.40	5	200.65	77.00
10	37.70	4	89.00	62.00

## ■ (2) Replace Missing Values with Mode or Mean

- ▶ Mode of categorical field cylinders = 4
- ▶ Missing values replaced with this value
- ▶ Mean for non-missing values in numeric field cubicinches = 200.65
- ▶ Missing values replaced with 200.65

# Handling Missing Data (cont'd)

- ▶ Substituting mode or mean for missing values sometimes works well
- ▶ Mean not always best choice for “typical” value
- ▶ Resulting confidence levels for statistical inference become overoptimistic (Larose)
- ▶ Domain experts should be consulted regarding approach to replace missing values
- ▶ Benefits and drawbacks resulting from the replacement of missing values must be carefully evaluated



# Handling Missing Data (*cont'd*)



	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.00	8	350.00	165.00
2	31.90	4	89.00	71.00
3	517.00	8	302.00	140.00
4	15.00	8	400.00	150.00
5	30.50	4	144.15	116.55
6	23.00	4	350.00	125.00
7	13.00	6	351.00	158.00
8	14.00	8	323.45	215.00
9	25.40	5	81.84	77.00
10	37.70	4	89.00	62.00

## ■ (3) Replace Missing Values with Random Values

- ▶ Values randomly taken from underlying distribution
- ▶ Value for cylinders, cubicinches, and hp randomly drawn proportionately from each field's distribution
- ▶ Method superior compared to mean substitution
- ▶ Measures of location and spread remain closer to original

# Handling Missing Data (cont'd)

- ▶ No guarantee resulting records make sense
- ▶ Suppose randomly-generated values `cylinders = 8` and `cubicinches = 82`
- ▶ A strange engine size?
- ▶ Alternate methods strive to replace values more precisely
- ▶ What is likely value, given record's other attribute values?
- ▶ For example, American car has 300 cubic inches and 150 horsepower
- ▶ Japanese car has 100 cubic inches and 90 horsepower
- ▶ American car expected to have more cylinders

# Identifying Misclassifications

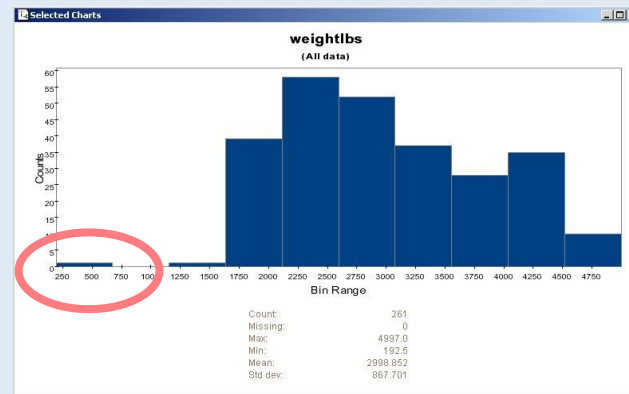
origin	
Level Name	Counts
USA	1
France	1
US	156
Europe	46
Japan	51

- ▶ Verify values valid and consistent
- ▶ Frequency distribution shows five classes: USA, France, US, Europe, and Japan
- ▶ Count for USA = 1 and France = 1?
- ▶ Two records classified inconsistently with respect to origin of the manufacturer
- ▶ Maintain consistency by labeling USA → US, and France → Europe

# Graphical Methods for Identifying Outliers

- ▶ Outliers are values that lie near extreme limits of data range
- ▶ Outliers may represent errors in data entry
- ▶ Certain statistical methods very sensitive to outliers and may produce unstable results
- ▶ Neural Networks and k-Means benefit from normalized data

# Graphical Methods for Identifying Outliers (cont'd)

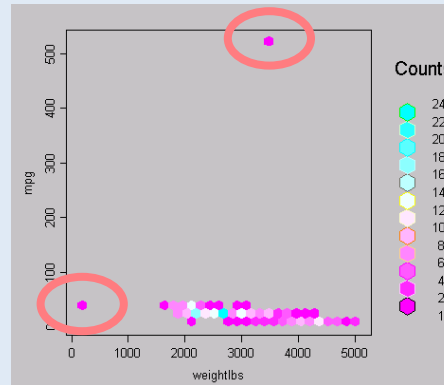


- ▶ A histogram examines values of numeric fields
- ▶ This histogram shows vehicle weights for cars data set
- ▶ The extreme left-tail contains one outlier weighing several hundred pounds (192.5)
- ▶ Should we doubt validity of this value?

# Graphical Methods for Identifying Outliers (cont'd)

- ▶ Analysis of weightlbs field shows remaining records contain whole-numbered (no decimal) values
- ▶ Perhaps value of 192.5 is an error
- ▶ Should it be 1925?
- ▶ Cannot know for sure and requires further investigation
- ▶ Discuss the meaning of the value with someone familiar with database content

# Graphical Methods for Identifying Outliers (cont'd)



- ▶ Two-dimensional scatter plots help determine outliers between variable pairs
- ▶ Scatter plot of mpg against weightlbs shows two possible outliers
- ▶ Most data points cluster together along x-axis
- ▶ However, one car weighs 192.5 pounds and other gets over 500 miles per gallon?

# Data Transformation

- ▶ Variables tend to have ranges different from each other
- ▶ In baseball, two fields may have ranges:
- ▶ Batting average: [ 0.0, 0.400 ]
- ▶ Number of home runs: [ 0, 70 ]
- ▶ Some data mining algorithms adversely affected by differences in variable ranges
- ▶ Variables with greater ranges tend to have larger influence on data model's results
- ▶ Therefore, numeric field values should be normalized
- ▶ Standardizes scale of effect each variable has on results



# Data Transformation (cont'd)

- **Two prevalent normalization techniques available**

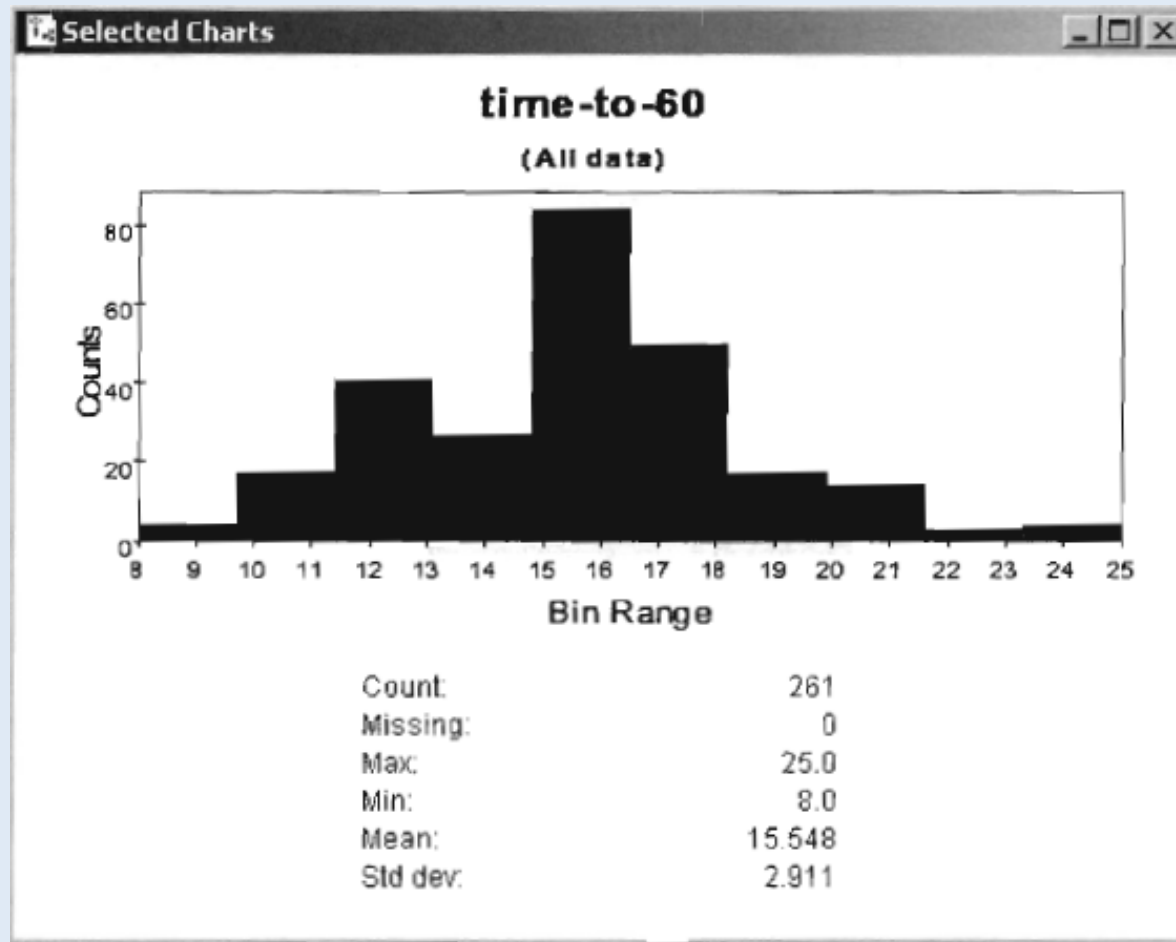
- **(1) Min-Max Normalization**

- ▶ Determines how much greater field value is than minimum value for field
- ▶ Scales this difference by field's range

$$X^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- ▶ From Figure 2.7, Min = 8 and Max = 25 for time-to-60 field

# Data Transformation (cont'd)



**Figure 2.7** Histogram of *time-to-60*, with summary statistics.

# Data Transformation (cont'd)

- ▶ Find Min-max Normalization for three automobiles having time-to-60 values 8, 15.548, and 25
- ▶ Example: “drag-racing-ready” is fastest vehicle and takes 8 seconds to reach 60mph
- ▶  $\text{Min}(\text{time-to-60}) = 8$
- ▶ Using Min-max Normalization:

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} = \frac{8 - 8}{25 - 8} = 0$$

- ▶ Minimum field values for time-to-60 have Min-max Normalization value = 0

# Data Transformation (cont'd)

- ▶ Example: Assume “average” vehicle takes 15.548 seconds to reach 60mph
- ▶  $\text{Avg}(\text{time-to-60}) = 15.548$
- ▶ Using Min-max Normalization:

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} = \frac{15.548 - 8}{25 - 8} = 0.444$$

- ▶ Field values near center of field's distribution have Min-max Normalization values near 0.5

# Data Transformation (cont'd)

- ▶ Example: The “I’ll get there when I’m ready” vehicle takes 25 seconds to reach 60mph
- ▶  $\text{Max}(\text{time-to-60}) = 25$
- ▶ Using Min-max Normalization:

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} = \frac{25 - 8}{25 - 8} = 1.0$$

- ▶ Maximum field values have Min-max Normalization value = 1
- ▶ Min-max Normalization values range [ 0, 1 ]

# Data Transformation (cont'd)

## ■ (2) Z-score Standardization

- ▶ Widely used in statistical analysis
- ▶ Takes difference between field value and field value mean
- ▶ Scales this difference by field's standard deviation

$$X^* = \frac{X - \text{mean}(X)}{\text{SD}(X)}$$

- ▶ From Figure 2.7, mean and standard deviation of time-to-60 equal 15.548 and 2.911, respectively

# Data Transformation (cont'd)

- ▶ Example: Again, consider “drag-racing-ready” automobile which takes 8 seconds to reach 60mph
- ▶ Using Z-score Standardization:

$$X^* = \frac{X - \text{mean}(X)}{\text{SD}(X)} = \frac{8 - 15.548}{2.911} = -2.593$$

- ▶ Data values that lie below the mean have negative Z-score Standardization values

# Data Transformation (cont'd)

- ▶ **Example: Our “average” vehicle takes exactly 15.548 seconds to reach 60mph**
- ▶ **Using Z-score Standardization:**

$$X^* = \frac{X - \text{mean}(X)}{\text{SD}(X)} = \frac{15.548 - 15.548}{2.911} = 0$$

- ▶ **Data values equal to field's mean have Z-score Standardization value = 0**



# Data Transformation (cont'd)

- ▶ Example: Slowest car, “I’ll get there when I’m ready”, takes 25 seconds to reach 60mph
- ▶ Using Z-score Standardization:

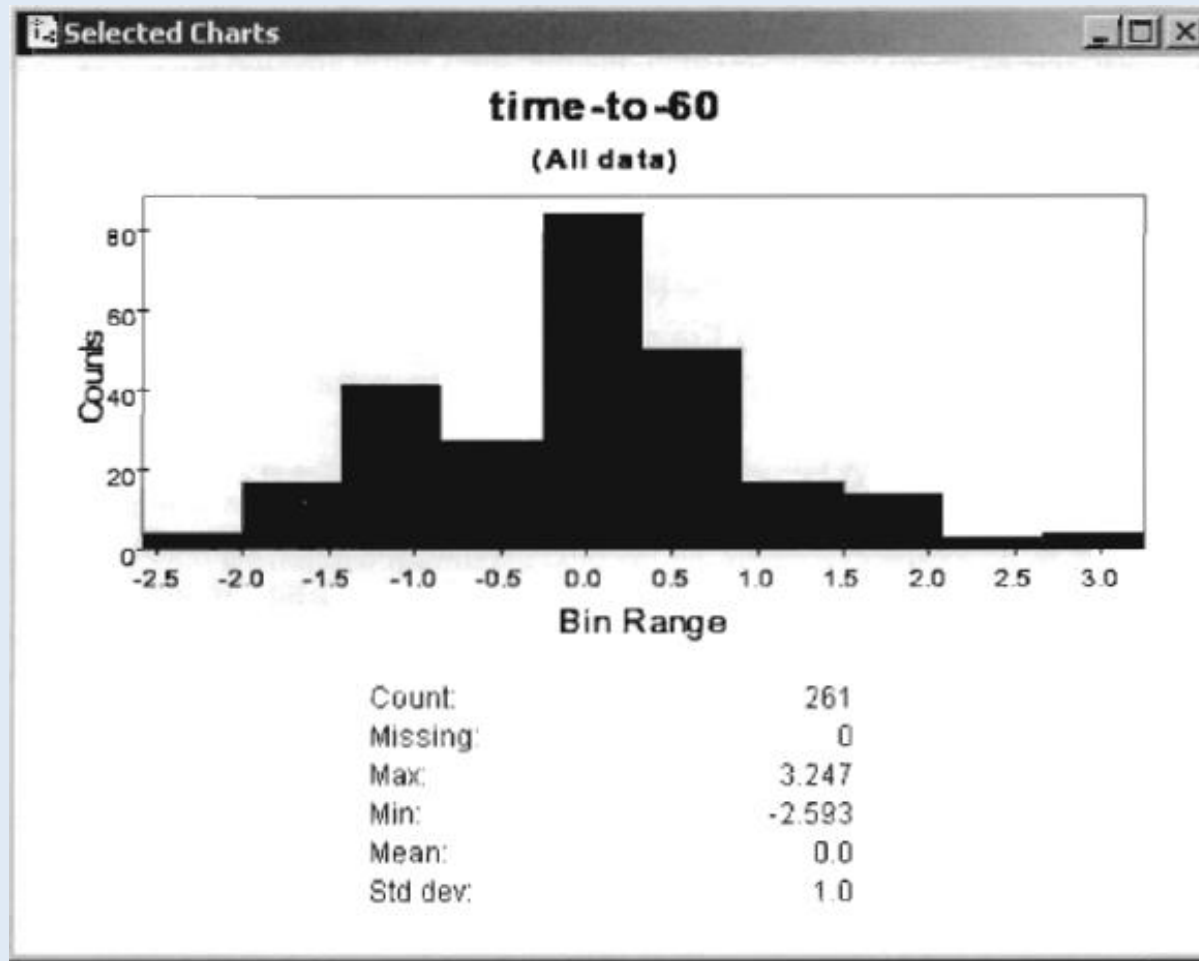
$$X^* = \frac{X - \text{mean}(X)}{\text{SD}(X)} = \frac{25 - 15.548}{2.911} = 3.247$$

- ▶ Data values that lie above the mean have positive Z-score Standardization values

# Data Transformation (cont'd)

- ▶ In summary, Z-score Standardization values typically range [ -4, 4 ]
- ▶ Field values below field mean → negative Z-score Standardization values
- ▶ Field values equal to field mean → Z-score Standardization value = 0
- ▶ Field values above field mean → positive Z-score Standardization values

# Data Transformation (cont'd)



**Figure 2.8** Histogram of *time-to-60* after Z-score standardization.

# Numerical Methods for Identifying Outliers

- ▶ Using Z-score Standardization to Identify Outliers
- ▶ Outliers are Z-score Standardization values either less than -3, or greater than 3
- ▶ Values outside range  $[-3, 3]$  require further investigation to determine their validity
- ▶ For example, the “I’ll get there when I’m ready” car has Z-score Standardization value = 3.247 (greater than 3)
- ▶ Identified as an outlier
- ▶ Validity of this value may be determined by discussing results with someone familiar with database content

# Numerical Methods for Identifying Outliers (cont'd)

- Unfortunately, mean and standard deviation used in Z-score Standardization formula are sensitive to outliers
- In other words, mechanism used to detect outliers is unduly affected by presence of outliers
- May be appropriate to choose method for detecting outliers not sensitive to their presence

# Numerical Methods for Identifying Outliers (cont'd)

- ▶ Using Interquartile Range (IQR) to Identify Outliers
- ▶ Robust statistical method and less sensitive to presence of outliers
- ▶ Data divided into four quartiles, each containing 25% of data
- ▶ First quartile (Q1) 25th percentile
- ▶ Second quartile (Q2) 50th percentile (median)
- ▶ Third quartile (Q3) 75th percentile
- ▶ Fourth quartile (Q4) 100th percentile
- ▶ IQR is measure of variability in data

# Numerical Methods for Identifying Outliers (cont'd)

- ▶  $IQR = Q3 - Q1$  and represents spread of middle 50% of the data
- ▶ Data value defined as outlier if located:
  - ▶  $1.5 \times (IQR)$  or more below  $Q1$ ; or
  - ▶  $1.5 \times (IQR)$  or more above  $Q3$
- ▶ For example, set of test scores have 25th percentile ( $Q1$ ) = 70, and 75th percentile ( $Q3$ ) = 80
- ▶ 50% of test scores fall between 70 and 80 and Interquartile Range ( $IQR$ ) =  $80 - 70 = 10$
- ▶ Test scores are identified as outliers if:
  - ▶ Lower than  $Q1 - 1.5 \times (IQR) = 70 - 1.5(10) = 55$ ; or
  - ▶ Higher than  $Q3 + 1.5 \times (IQR) = 80 + 1.5(10) = 95$