Discovering Knowledge in Data

Daniel T. Larose, Ph.D.

Chapter 2
Data Preprocessing

Prepared by James Steck and Eric Flores

CRISP-DM Review

Business Research Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment

- This chapter (chapter 2) examines phases 2 and 3 of the CRISP-DM process
- Chapter 3 expands on the Data Understanding phase
- Chapter 4 and above focus on Modeling

Why Do We Preprocess Data?

- Raw data is often unprocessed, incomplete, or noisy.
- May contain:
  - Obsolete/redundant fields
  - Missing values
  - Outliers
  - Data in form not suitable for data mining
  - Values not consistent with policy or common sense

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

- For data mining purposes, database values must undergo <u>data cleaning</u> and <u>data transformation</u>
- Data often from legacy (out-dated) databases where values:
  - Not looked at in years
  - Expired
  - No longer relevant
  - Missing
- Minimize GIGO (Garbage In → Garbage Out)
  - IF <u>**G**</u>arbage <u>**I**</u>nto model is <u>minimized</u> →
    THEN <u>**G**</u>arbage results <u>**O**</u>ut from model is <u>minimized</u>
- Effort for Data preparation ranges around10%-60% of data mining process – depending on dataset

Data Cleaning – Example

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

CustomerID field is assumed to be fine; But Zip Code, Gender?

- Zip Code
  - Do not assume local format
    - 90210 (U.S.) vs. J2S7K7 (Canada)
    - In a free trade era should expect some unusual values
  - Be aware of data type/conversion issues
    - Zip code 06269 stored in numeric field truncates the leading zeroes, and thus, is represented as 6269 (Zip Code for Storrs, CT)
- Gender
  - Value is missing for customer 1003

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

Data Cleaning – Example *(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

- Income Field Contains $10,000,000?
  - Possibly valid on zip code 90210 (Beverly Hills, CA)
  - Still considered outlier (extreme data value) - Some statistical and data mining methods negatively affected by outliers
  - Handling of outliers examined later in this chapter
- 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
- Income Field Contains $99,999?
  - Value may be valid, but…other values appear rounded to nearest $5,000
  - Legacy Systems: Value represents database code used to denote missing value?
- Other considerations for Income
  - 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?

Data Cleaning – Example *(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

- Age field contains C
  - Possible a leftover of earlier categorization of age into a bin labeled C?
  - Data Mining software will likely reject a text value on an otherwise numeric field – this needs resolution
- Age field contains 0 (zero)
  - Unlikely: A newborn baby made $1000 transaction
  - Most probably: Missing value or other anomalous condition coded as 0 (zero)
  - **Important**: Age value will quickly become obsolete; it is recommended to store date type fields (like birthdate) instead, and calculate age as needed
- Marital Status Field
  - What is the meaning of the symbols?
  - Don't make assumptions: Is S for Single or Separated?
  - Consider possibility of codes using words from another language: C is for Cold in English, and Chaud (Hot) in French
- Transaction Amount Field
  - Values in this fields seems OK, assuming common unit of measure

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

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, all things being equal, having more data is almost always better
- Careful analysis required to handle issue

Handling Missing Data *(cont'd)*

- Suppose you are given a *cars* dataset containing records for 261 automobiles manufactured in 1970s and 1980s
- Suppose that some fields are missing for certain records, like in figure below:

	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900		71	Europe
3	17.000	302	140	US
4	15.000	400	150	
5	37.700	89	62	Japan

- Delete Records Containing Missing Values?
  - Dangerous, as pattern of missing values may be systematic
  - Valuable information in other fields lost
    - As much as 80% of the records lost, if only 5% of data values are missing, according to Schmueli, Patel, and Bruce [1].
- Three alternative methods available – Not entirely satisfactory
- Data imputation methods – Better approach

Handling Missing Data *(cont'd)*

- Alternative Method #1- Replace Missing Values with a Constant, specified by the Analyst
- Example:
  - Missing numeric values are replaced with 0.0
  - Missing categorical values are replaced with "Missing"

	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900	0	71	Europe
3	17.000	302	140	US
4	15.000	400	150	Missing
5	37.700	89	62	Japan

Handling Missing Data *(cont'd)*

- Alternative Method #2 - Replace Missing Values with Mode or Mean
- Example:
  - Mode of categorical field *brand* = US
    - Missing values are replaced with this value
  - Mean for non-missing values in numeric field *cubicinches* = 200.65
    - Missing values are replaced with 200.65

	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900	200.65	71	Europe
3	17.000	302	140	US
4	15.000	400	150	US
5	37.700	89	62	Japan

Handling Missing Data *(cont'd)*

- Notes on Alternative Method #2 - Replace Missing Values with <u>Mode</u> or <u>Mean</u>
  - Substituting mode or mean for missing values sometimes works well – however, end user needs to be informed.
  - Mean is not always the best choice for "typical" value.
    - For example, the mean maybe greater than 80-th percentile.
  - Resulting confidence levels for statistical inference become overoptimistic (Larose), since measures of spread are artificially reduced.
  - Benefits and drawbacks resulting from the replacement of missing values must be carefully evaluated against possible invalidity of results.

Handling Missing Data *(cont'd)*

- Alternative Method #3 - Replace Missing Values with Random Values
  - Example: Value for *cylinders*, *cubicinches*, and *hp* randomly drawn proportionately from each field's distribution
  - Values randomly taken from underlying distribution
  - Benefit: Measures of location and spread remain closer to original
  - No guarantee that resulting records would make sense (see side note)

This record leads to a car that does not exist!

Japanese car with 400cc engine

	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900	450	71	Europe
3	17.000	302	140	US
4	15.000	400	150	Japan
5	37.700	89	62	Japan

Handling Missing Data *(cont'd)*

- Data Imputation Methods
  - Imputation of Missing Data - What is the likely value, given record's other attribute values?
  - Example: From two samples below, American car would be expected to have more cylinders
    - American car with 300 cubic inches and 150 horsepower
    - Japanese car with 100 cubic inches and 90 horsepower
  - Requires tools like multiple regression, or classification and regression trees
  - To be discussed in Chapter 13 – *Imputation of Missing Data*

Identifying Misclassifications

- Check classification labels to verify values are valid and consistent
- Example: Table below – Frequency distribution for origin of manufacture of automobiles
  - Frequency distribution shows four classes: USA, France, US, and Europe.
  - Count for USA = 1 and France = 1.
  - Two records classified inconsistently with respect to origin of the manufacture.
  - Maintain consistency by labeling USA → US, and France → Europe.

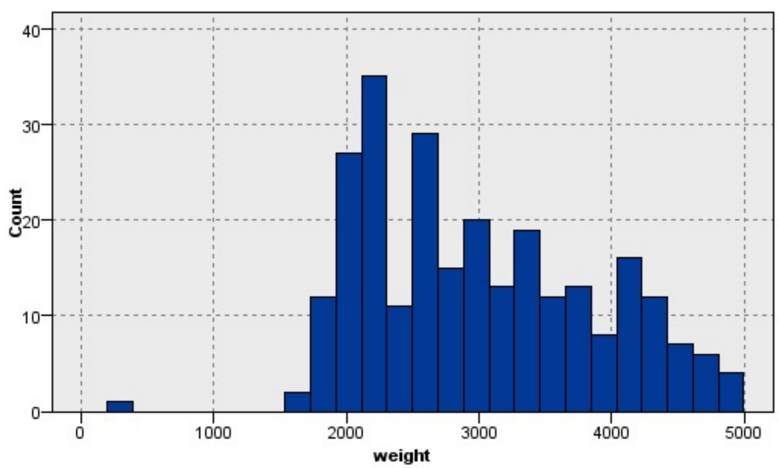
Brand	Frequency
USA	1
France	1
US	156
Europe	46

Graphical Methods for Identifying Outliers

- Outliers are extreme values that go against the trend of the remaining data
- Outliers may represent errors in data entry
- Even if valid data point, certain statistical methods are very sensitive to outliers and may produce unstable results
- Two graphical methods presented

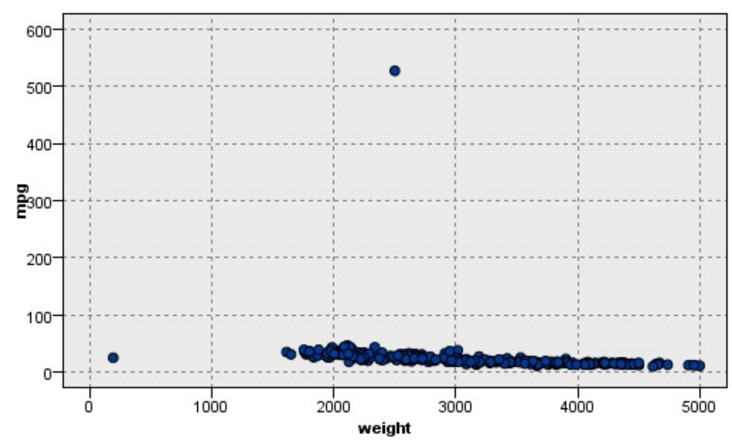
Graphical Methods for Identifying Outliers *(cont'd)*

- Method #1 - Histogram
  - A <u>histogram</u> examines values of <u>numeric</u> fields (good for one-dimensional data)
  - Example: Histogram shows vehicle weights for a *cars* data set
    - The extreme left-tail contains one outlier weighing several hundred pounds (192.5)
    - Should we doubt validity of this value? This is too light for a car.
    - Possibility: Original value was 1925 pounds. Requires further investigation.



Graphical Methods for Identifying Outliers *(cont'd)*

- Method #2 – Two (or three)-dimensional Scatter Plot
  - Two (or three)-dimensional scatter plots help determine outliers in two (or three) dimensions.
  - Example: Scatter plot of *mpg* against *weight (lbs)* 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!
    - Important: A record may be outlier in a particular dimension, but not in the other



Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

Measures of Center and Spread

Measures of center (1/5) - Introduction

- Estimate where the center of a particular variable's distribution lies
- Most common *measures of center*
  - Mean, Median and Mode
    - They are a special case of *measures of location*, which indicate where a numeric variables lies.

Measures of Center and Spread (cont'd)

Measures of center (2/5) - Mean

- Average of the valid values for a random variable
  - Add all field values and divide by sample size
  - Denoted as $\bar{x}$ (x-bar) and computed as:

$$\bar{x} = \frac{\sum x}{n}$$

    - Where
      - $\sum$ represents "sum of all variables"
      - n represents sample size

Measures of Center and Spread *(cont'd)*

Measures of center (3/5) - Example

- From the table below, use the Sum and Count to calculate the Mean

$$\bar{x} = \frac{\sum x}{n} = \frac{5209}{3333} = 1.563$$

Population: Number of calls made by each customer

Customer Service Calls

Statistics

Count	3333
Mean	1.563
Sum	5209.000
Median	1
Mode	1

Measures of Center and Spread *(cont'd)*

Measures of center (4/5) – Alternatives

- Mean is not always ideal
  - On extremely skewed datasets, it is less representative of variable center; it is also sensitive to outliers
- Alternative measures of center
  - Median – Field value in the middle, when field values are sorted into ascending order
  - Mode – Field value occurring with the greatest frequency
    - Pros: Can be used with either numerical or categorical data
    - Cons: Not always associated with the variable center

Measures of Center and Spread (cont'd)

Measures of center (5/5) – Further notes

- Measures of center do not always concur
- Example: Table below
  - Median is 1 – Half the customers made one customer service call
  - Mode is 1 - Most frequent number of calls is one
  - But Mean is 1.563 – (56.3% higher than median/mode) Caused by the mean sensitivity to the right-skewness of the data

Population: Number of calls made by each customer

Customer Service Calls

Statistics

Count	3333
Mean	1.563
Sum	5209.000
Median	1
Mode	1

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

Measures of Center and Spread

Measures of Spread (1/5) - Introduction

- Measures of location not enough to summarize a variable
- Example: Table with Price/Earning (P/E) ratios for two portfolios
  - Portfolio A – Spread with one very low and one very high P/E value
  - Portfolio B – Tightly clustered around the center
  - P/E ratios for each portfolio is distinctly different, yet **they both** have P/E ratios with mean 10, median 11 and mode 11.
- Clearly, measures of center do not provide a complete picture
- Measures of spread or measures of variability complete the picture by describing how spread the data values of each portfolio are

Stock Portfolio A	Stock Portfolio B
1	7
11	8
11	11
11	11
16	13

P/E ratio for the first stock

Measures of Center and Spread

Measures of Spread (2/5) - Introduction

- Typical measures of variability include
  - Range (maximum – minimum)
  - Standard Deviation – Sensitive to the presence of outliers (because of the squaring (power 2) involved – see below)
  - Mean Absolute Deviation – Preferred in situations involving extreme values
- Sample Standard Deviation is defined by

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

  - Interpreted as "typical" distance between a field value and the mean
  - Most field values (95%) lie within two standard deviations of the mean
    - Example: For table below, number of calls made by most customers are within 2(1.315) = 2.63 of the mean of 1.563 calls. Most customers made between -1.067 and 4.193 (rounded to integers 0 to 4) calls.

Customer Service Calls

Statistics

Count	3333
Mean	1.563
Sum	5209.000
Median	1
Mode	1

Population: Number of calls made by each customer

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 models' results
- Therefore, <u>numeric</u> field values should be <u>normalized</u>
- Standardizing will scale the effect each variable has on results
- Neural Networks and other algorithms that make use of distance measures benefit from normalization
- Two of the prevalent methods will be reviewed
- In the following pages X* will refer to the normalized form of random variable X

Min-Max Normalization

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

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

- Figure 2.8 below shows summary statistics for *weight (lbs)* field
  - Min = 1613
  - Max = 4997

weightlbs

Statistics

Mean	3005.490
Min	1613
Max	4997
Range	3384
Standard Deviation	852.646

Min-Max Normalization *(cont'd)*

Find Min-Max normalization for cars weighing 1613, 3305 and 4997 pounds, respectively

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

Where:

min(X) = 1613

max(X) = 4997

Car	Weightlbs	Formula	Result	Comments
Lightest vehicle	X = 1613	$X^* = \frac{1613 - 1613}{4997 - 1613}$	X* = 0	Represents the minimum value in this variable, and has min-max normalization of zero.
Mid-range vehicle	X = 3305	$X = \frac{3305 - 1613}{4997 - 1613}$	X* = 0.5	Weight exactly half-weight between the lightest and the heaviest vehicle, and has min-max normalization of 0.5.
Heaviest vehicle	X = 4997	$X = \frac{4997 - 1613}{4997 - 1613}$	X* = 1	Heaviest vehicle of the dataset has min-max normalization of one.

Min-Max normalization will always have a value between 0 and 1.

It is also possible to find the associated data value for a given Min-Max Normalization (how?)

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

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

- Figure 2.8 below shows that mean (weight) and standard deviation for weight equals 3005.49 and 852.646, respectively

weightlbs

Statistics

Mean	3005.490
Min	1613
Max	4997
Range	3384
Standard Deviation	852.646

Z-score Standardization *(cont'd)*

Find Z-scode standardization for cars weighing 1613, 3006 and 4997 pounds, respectively

Where:

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

mean(X) = 3005.49

SD(X) = 852.65

Car	Weightlbs	Formula	Result	Comments
Lightest vehicle	X = 1613	$X^* = \frac{1613 - 3005.49}{852.646}$	X* ≈ -1.63	Data values below the mean will have negative Z-score standardization.
Average vehicle	X = 3005.49	$X = \frac{3005.49 - 3005.49}{852.646}$	X* ≈ 0	Values falling very close to the mean will have zero (0) Z-score
Heaviest vehicle	X = 4997	$X = \frac{4997 - 3005.49}{852.646}$	X* ≈ 2.34	Data values above the mean will have a positive Z-score standardization

It is also possible to find the associated data value for a given Z-score (how?).