

Lecture 2 - Data Preprocessing

Adnan Ferdous Ashrafi

Stamford University Bangladesh



Table of Contents

1 Data Preprocessing

- Definition

2 WHY DO WE NEED TO PREPROCESS THE DATA?

- Reasoning
- Objective

3 Data Cleaning

- Definition
- An Example

4 HANDLING MISSING DATA

- Definition
- An Example
- Missing data handling techniques

5 IDENTIFYING MISCLASSIFICATIONS

- Definition
- Removing misclassifications

6 GRAPHICAL METHODS FOR IDENTIFYING OUTLIERS

- Definition of Outliers
- Histograms
- Scatterplots

7 Data Transformation

- Definition
- Min–Max Normalization
- Z-Score Standardization

8 NUMERICAL METHODS FOR IDENTIFYING OUTLIERS

- Definition
- Interquartile range
- Outlier detection

9 Getting to know your data

- Suggestions

10 Further Reading

Data Preprocessing

Definition

Data preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance, and is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: -100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), and missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running any analysis. [1]

WHY DO WE NEED TO PREPROCESS THE DATA?

Much of the raw data contained in databases is unpreprocessed, incomplete, and noisy. For example, the databases may contain:

- Fields that are obsolete or redundant
- Missing values
- Outliers
- Data in a form not suitable for data mining models
- Values not consistent with policy or common sense.

To be useful for data mining purposes, the databases need to undergo preprocessing, in the form of data cleaning and data transformation. Data mining often deals with data that hasn't been looked at for years, so that much of the data contains field values that have expired, are no longer relevant, or are simply missing.

Objectives

The overriding objective is to minimize GIGO: to minimize the “garbage” that gets into our model so that we can minimize the amount of garbage that our models give out.

Dorian Pyle, in his book *Data Preparation for Data Mining* [1], estimates that data preparation alone accounts for 60% of all the time and effort expended in the entire data mining process. In this chapter we examine two principal methods for preparing the data to be mined, data cleaning, and data transformation.

Data Cleaning

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. [2] Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting.

Example data that needs cleaning

To illustrate the need to clean up data, let's take a look at some of the types of errors that could creep into even a tiny data set, such as that in Table 1.

Table 1: Can You Find Any Problems in This Tiny Data Set?

Customer ID	ZIP	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	7500	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	M	99999	30	D	3000

Example data that needs cleaning - ZIP Code

Let's assume that we are expecting all of the customers in the database to have the usual five-numeral U.S. zip code. Now, customer 1002 has this strange (to American eyes) zip code of J2S7K7. If we were not careful, we might be tempted to classify this unusual value as an error and toss it out, until we stop to think that not all countries use the same zip code format. Actually, this is the zip code of St. Hyancinthe, Quebec, Canada, so probably represents real data from a real customer. What has evidently occurred is that a French-Canadian customer has made a purchase and put their home zip code down in the field required. Especially in this era of the North American Free Trade Agreement, we must be ready to expect unusual values in fields such as zip codes, which vary from country to country.

Example data that needs cleaning - ZIP Code

What about the zip code for customer 1004? We are unaware of any countries that have four-digit zip codes, such as the 6269 indicated here, so this must be an error, right? Probably not. Zip codes for the New England states begin with the numeral 0. Unless the zip code field is defined to be character (text) and not numeric, the software will probably chop off the leading zero, which is apparently what happened here. The zip code is probably 06269, which refers to Storrs, Connecticut, home of the University of Connecticut.

Example data that needs cleaning - Gender

The next field, gender, contains a missing value for customer 1003. We detail methods for dealing with missing values later in the chapter. The income field, which we assume is measuring annual gross income, has three potentially anomalous values. First, customer 1003 is shown as having an income of \$10,000,000 per year. Although entirely possible, especially when considering the customer's zip code (90210, Beverly Hills), this value of income is nevertheless an outlier, an extreme data value. Certain statistical and data mining modeling techniques do not function smoothly in the presence of outliers; we examine methods of handling outliers later in the chapter.

Example data that needs cleaning - Income

Poverty is one thing, but it is rare to find an income that is negative, as our poor customer 1004 has. Unlike customer 1003's income, customer 1004's reported income of -\$40,000 lies beyond the field bounds for income and therefore must be an error. It is unclear how this error crept in, with perhaps the most likely explanation being that the negative sign is a stray data entry error. However, we cannot be sure and should approach this value cautiously, attempting to communicate with the database manager most familiar with the database history.

Example data that needs cleaning - Income

So what is wrong with customer 1005's income of \$99,999? Perhaps nothing; it may in fact be valid. But if all the other incomes are rounded to the nearest \$5000, why the precision with customer 1005? Often, in legacy databases, certain specified values are meant to be codes for anomalous entries, such as missing values. Perhaps 99999 was coded in an old database to mean missing. Again, we cannot be sure and should again refer to the “wetware.”

Example data that needs cleaning - Income

Finally, are we clear as to which unit of measure the income variable is measured in? Databases often get merged, sometimes without bothering to check whether such merges are entirely appropriate for all fields. For example, it is quite possible that customer 1002, with the Canadian zip code, has an income measured in Canadian dollars, not U.S. dollars.

Example data that needs cleaning - Age

The age field has a couple of problems. Although all the other customers have numerical values for age, customer 1001's "age" of C probably reflects an earlier categorization of this man's age into a bin labeled C. The data mining software will definitely not like this categorical value in an otherwise numerical field, and we will have to resolve this problem somehow. How about customer 1004's age of 0? Perhaps there is a newborn male living in Storrs, Connecticut, who has made a transaction of \$1000. More likely, the age of this person is probably missing and was coded as 0 to indicate this or some other anomalous condition (e.g., refused to provide the age information).

Example data that needs cleaning - Age

Of course, keeping an age field in a database is a minefield in itself, since the passage of time will quickly make the field values obsolete and misleading. It is better to keep date-type fields (such as birthdate) in a database, since these are constant and may be transformed into ages when needed.

Example data that needs cleaning - Marital Status

The marital status field seems fine, right? Maybe not. The problem lies in the meaning behind these symbols. We all think we know what these symbols mean, but are sometimes surprised. For example, if you are in search of cold water in a rest room in Montreal and turn on the faucet marked C, you may be in for a surprise, since the C stands for chaud, which is French for hot. There is also the problem of ambiguity. In Table 1, for example, does the S for customers 1003 and 1004 stand for single or separated?

HANDLING MISSING DATA

Missing data is a problem that continues to plague data analysis methods. Even as our analysis methods gain sophistication, we continue to encounter missing values in fields, especially in databases with a large number of fields. The absence of information is rarely beneficial. All things being equal, more data is almost always better. Therefore, we should think carefully about how we handle the thorny issue of missing data.

Example Data that is missing

We will introduce ourselves to a new data set, the cars data set, originally compiled by Barry Becker and Ronny Kohavi of Silicon Graphics. The data set consists of information about 261 automobiles manufactured in the 1970s and 1980s, including gas mileage, number of cylinders, cubic inches, horsepower, and so on.

	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.0	8	350.0	165.0
2	31.9	4	89.0	71.0
3	517.0	8	302.0	140.0
4	15.0		400.0	150.0
5	30.5			
6	23.0		350.0	125.0
7	13.0		351.0	158.0
8	14.0	8		215.0
9	25.4	5		77.0
10	37.7	4	89.0	62.0

Figure 1: Some of our field values are missing!

Techniques in handling missing data

A common method of handling missing values is simply to omit from the analysis the records or fields with missing values. However, this may be dangerous, since the pattern of missing values may in fact be systematic, and simply deleting records with missing values would lead to a biased subset of the data. Further, it seems like a waste to omit the information in all the other fields, just because one field value is missing. Therefore, data analysts have turned to methods that would replace the missing value with a value substituted according to various criteria.

- ① Replace the missing value with some constant, specified by the analyst.
- ② Replace the missing value with the field mean (for numerical variables) or the median (for categorical variables).
- ③ Replace the missing values with a value generated at random from the variable distribution observed.

Technique - replace with constant

Table View (17)

Rounding Tools

	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

Figure 2: Replacing missing field values with user-defined constants.

Technique - replace with mean/median

Table View (17)

Rounding Tools

	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

Figure 3: Replacing missing field values with means or medians.

Technique - replace with constant

Table View (17)

Rounding Tools

	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

Figure 4: Replacing missing field values with random draws from the distribution of the variable.

Identifying misclassifications

Let us look at an example of checking the classification labels on the categorical variables, to make sure that they are all valid and consistent. One of the useful functions for missing values node is to display a frequency distribution of the categorical variables available. For example, in Table 2:

Table 2: Notice Anything Strange about This Frequency Distribution?

Label Name	Count
USA	1
France	1
US	156
Europe	46
Japan	51

Removing misclassifications

The frequency distribution shows five classes: USA, France, US, Europe, and Japan. However, two of the classes, USA and France, have a count of only one automobile each. What is clearly happening here is that two of the records have been classified inconsistently with respect to the origin of manufacture. To maintain consistency with the remainder of the data set, the record with origin USA should have been labeled US, and the record with origin France should have been labeled Europe.

Definition of outliers

Definition

Outliers are extreme values that lie near the limits of the data range or go against the trend of the remaining data. Identifying outliers is important because they may represent errors in data entry. Also, even if an outlier is a valid data point and not in error, certain statistical methods are sensitive to the presence of outliers and may deliver unstable results. Neural networks benefit from normalization, as do algorithms that make use of distance measures, such as the k-nearest neighbor algorithm.

Histograms

One graphical method for identifying outliers for numeric variables is to examine a histogram of the variable. Figure 5 shows a histogram generated of the vehicle weights from the cars data set. There appears to be one lonely vehicle in the extreme left tail of the distribution, with a vehicle weight in the hundreds of pounds rather than in the thousands.

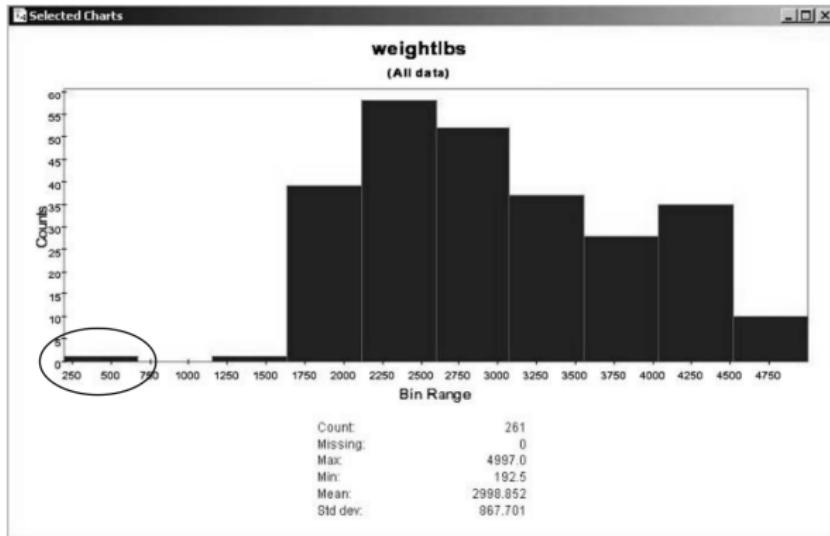


Figure 5: Histogram of vehicle weights: can you find the outlier?

Scatterplots

Sometimes two-dimensional scatter plots can help to reveal outliers in more than one variable. The scatter plot of mpg against weightlbs shown in Figure 6 seems to have netted two outliers.

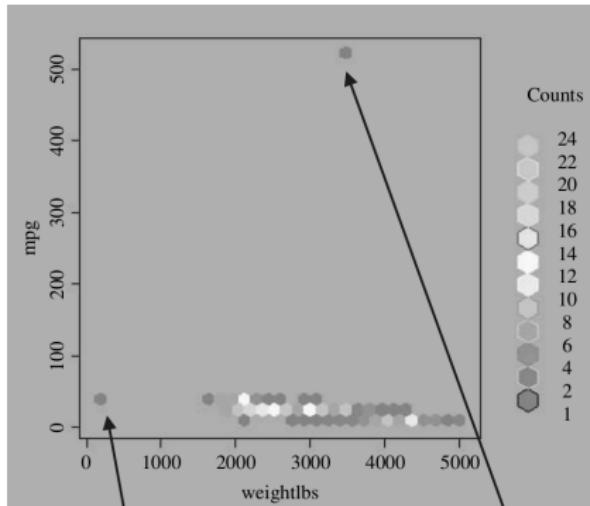


Figure 6: Scatter plot of mpg against weightlbs shows two outliers.

Data Transformation

Definition

Variables tend to have ranges that vary greatly from each other. For example, if we are interested in major league baseball, players' batting averages will range from zero to less than 0.400, while the number of home runs hit in a season will range from zero to around 70. For some data mining algorithms, such differences in the ranges will lead to a tendency for the variable with greater range to have undue influence on the results.

Therefore, data miners should normalize their numerical variables, to standardize the scale of effect each variable has on the results. There are several techniques for normalization, and we shall examine two of the more prevalent methods. Let X refer to our original field value and X^* refer to the normalized field value.

Min–Max Normalization

Min–max normalization works by seeing how much greater the field value is than the minimum value $\min(X)$ and scaling this difference by the range. That is,

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

For example, consider the time-to-60 variable from the cars data set, which measures how long (in seconds) each automobile takes to reach 60 miles per hour. Let's find the min–max normalization for three automobiles having times-to-60 of 8, 15.548, seconds, and 25 seconds, respectively.

Min-Max Normalization

Refer to Figure 7, a histogram of the variable time-to-60, along with some summary statistics.

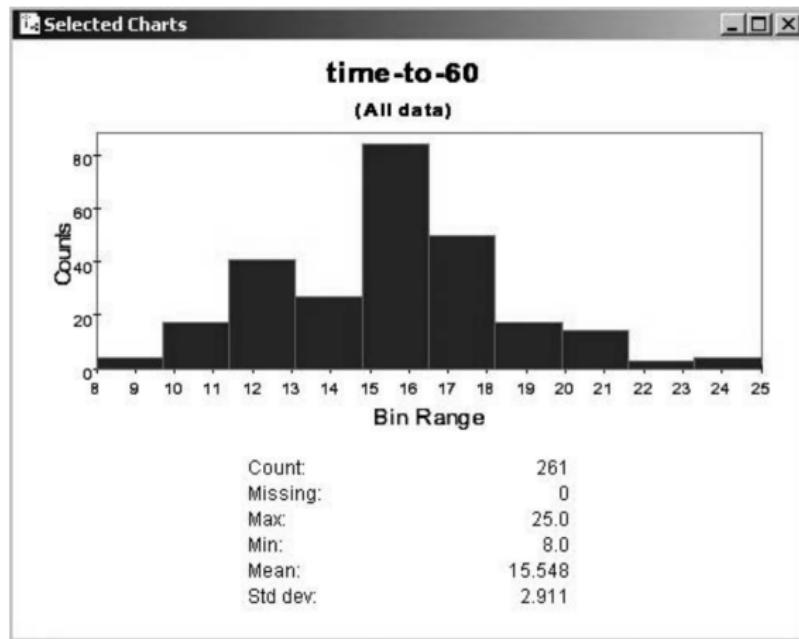


Figure 7: Histogram of time-to-60, with summary statistics.

Min–Max Normalization

- For a “drag-racing-ready” vehicle, which takes only 8 seconds (the field minimum) to reach 60 mph, the min–max normalization is

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

- For an “average” vehicle (if any), which takes exactly 15.548 seconds (the variable average) to reach 60 mph, the min–max normalization is

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

- For an “I’ll get there when I’m ready” vehicle, which takes 25 seconds (the variable maximum) to reach 60 mph, the min–max normalization is

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

Z-Score Standardization

Z-score standardization, which is very widespread in the world of statistical analysis, works by taking the difference between the field value and the field mean value and scaling this difference by the standard deviation of the field values. That is,

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

For example, consider the time-to-60 variable from the cars data set, which measures how long (in seconds) each automobile takes to reach 60 miles per hour. Let's find the min–max normalization for three automobiles having times-to-60 of 8, 15.548, seconds, and 25 seconds, respectively.

Z-Score Standardization

- For the vehicle that takes only 8 seconds to reach 60 mph, the Z-score standardization is:

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

- For an “average” vehicle (if any), which takes exactly 15.548 seconds (the variable average) to reach 60 mph, the Z-score standardization is

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

- For the car that takes 25 seconds to reach 60 mph, the Z-score standardization is

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

Z-Score Standardization

To summarize, Z-score standardization values will usually range between -4 and 4, with the mean value having a Z-score standardization of zero. Figure 8 is a histogram of the time-to-60 variable after the Z-score standardization of each field value. Note that the distribution is centered about zero and that the minimum and maximum agree with what we found above.

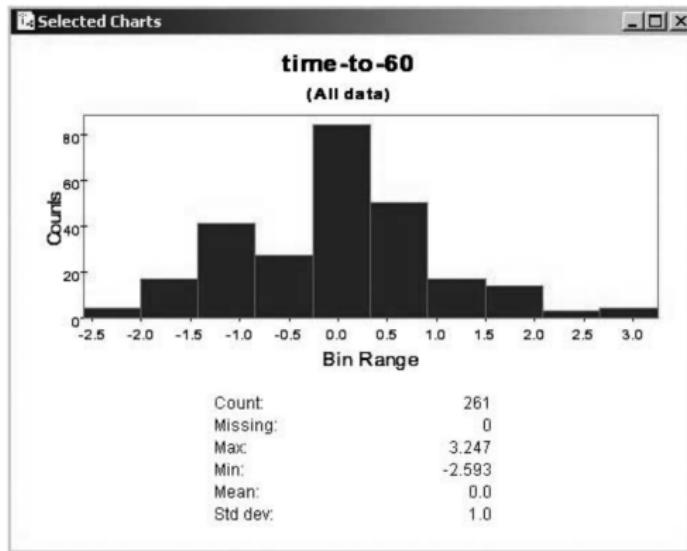


Figure 8: Histogram of time-to-60 after Z-score standardization.

NUMERICAL METHODS FOR IDENTIFYING OUTLIERS

One method of using statistics to identify outliers is to use Z-score standardization. Often, an outlier can be identified because it is much farther than 3 standard deviations from the mean and therefore has a Z-score standardization that is either less than -3 or greater than 3. Field values with Z-scores much beyond this range probably bear further investigation to verify that they do not represent data entry errors or other issues. For example, the vehicle that takes its time (25 seconds) getting to 60 mph had a Z-score of 3.247. This value is greater than 3 (although not by much), and therefore this vehicle is identified by this method as an outlier. The data analyst may wish to investigate the validity of this data value or at least suggest that the vehicle get a tune-up!

Interquartile range

The quartiles of a data set divide the data set into four parts, each containing 25% of the data.

- The first quartile (Q1) is the 25th percentile.
- The second quartile (Q2) is the 50th percentile, that is, the median.
- The third quartile (Q3) is the 75th percentile.

The interquartile range (IQR) is a measure of variability that is much more robust than the standard deviation. The IQR is calculated as $IQR = Q3 - Q1$ and may be interpreted to represent the spread of the middle 50% of the data.

Outlier detection

A robust measure of outlier detection is therefore defined as follows. A data value is an outlier if:

- a It is located $1.5(\text{IQR})$ or more below Q1 , or
- b It is located $1.5(\text{IQR})$ or more above Q3 .

Outlier detection - Example

For example, suppose that for a set of test scores, the 25th percentile was $Q1 = 70$ and the 75th percentile was $Q3 = 80$, so that half of all the test scores fell between 70 and 80. Then the interquartile range, the difference between these quartiles, was $IQR = 80 - 70 = 10$.

A test score would be robustly identified as an outlier if:

- a It is lower than $Q1 - 1.5(IQR) = 70 - 1.5(10) = 55$, or
- b It is higher than $Q3 + 1.5(IQR) = 80 + 1.5(10) = 95$.

Getting to know your data

There is no substitute for getting to know your data. Simple tools that show histograms of the distribution of values of nominal attributes, and graphs of the values of numeric attributes (perhaps sorted or simply graphed against instance number), are very helpful. These graphical visualizations of the data make it easy to identify outliers, which may well represent errors in the data file.

Data cleaning is a time-consuming and labor-intensive procedure but one that is absolutely necessary for successful data mining. With a large dataset, people often give up—how can they possibly check it all? Instead, you should sample a few instances and examine them carefully. You'll be surprised at what you find. Time looking at your data is always well spent.

Further Reading

- Chapter 2 of [Data Mining - Practical Machine Learning Tools and Techniques, Second Edition](#) - Ian H. Witten, Eibe Frank
- Chapter 2 of [DISCOVERING KNOWLEDGE IN DATA - An Introduction to Data Mining](#) - DANIEL T. LAROSE
- Chapter 2 of [Introduction to Data Mining \(Second Edition\)](#) - Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar

References

- [1] D. Pyle, *Data Preparation for Data Mining*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999.
- [2] S. Wu, “A review on coarse warranty data and analysis,” *Reliability Engineering & System Safety*, vol. 114, pp. 1–11, 2013. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0951832013000100>

*Thank you.
Any Questions?*