Solutions to Chapter 2
DATA PREPROCESSING

Prepared by James Cunningham, Graduate Assistant

3. **Explain why zip codes should be considered text variables rather than numeric.**

Zip codes should be considered text variables because they cannot be quantified on any numeric scale. Even their order has no numerical significance.

6. **True or false: All things being equal, more information is almost always better.**

The answer is true. In general, more information is almost always better. The more information we have to work with, the more insight into the underlying relationships of a particular domain of discourse we can glean from it.

7. **Explain why it is not recommended, as a strategy for dealing with missing data, to simply omit the records or fields with missing values from the analysis.**

It is not recommended to omit records or fields from an analysis simply because they have missing values. The rationale for this recommendation is that omitting these fields and records may cause us to lose valuable insight into the underlying relationships that we may have gleaned from the partial information that we do have.

8. **Which of the four methods for handling missing data would tend to lead to an underestimate of the spread (e.g., standard deviation) of the variable? What are some benefits to this method?**

Replacing a missing value by the attribute value's mean artificially reduces the measure of spread for that particular attribute. Although the mean value is not necessarily a typical value, for some data sets this form of substitution may work well. Specifically, the effectiveness of this technique depends on the size of the variation of the underlying population. In other words, the technique works well for populations having small variations, and works less effectively for populations having larger variations.

Several benefits to leveraging this method include (1) ease of implementation (i.e. only one value to impute), (2) preservation of the standard error (i.e. no additional residual error is introduced).

9. What are some of the benefits and drawbacks for the method for handling missing data that chooses values at random from the variable distribution?

By using the data values randomly generated from the variable distribution, the measures of center and spread are most likely to remain similar to the original; however, there is a chance that the resulting records may not make intuitive sense.

10. Of the four methods for handling missing data, which method is preferred?

Having the analyst choose a constant to replace missing values based on specific domain knowledge is overall, probably the most conservative choice. If missing values are replaced with a flag such as “missing” or “unknown”, in many situations those records would ultimately be excluded from the modeling process; that is, all remaining valid, potentially important, values contained in those records would not be included in the data model.

12. Make up a data set, consisting of the heights and weights of six children, in which one of the children is an outlier with respect to one of the variables, but not the other. Then alter this data set so that the child is an outlier with respect to both variables.

In the table below, Child #1 is an outlier with respect to Weight only. All children in the table are close in Height differing at most by 9 inches. However, all children except for Child # 1 are close in Weight differing at most by 7 pounds. Child #1 is an outlier as the Weight differs by 18 pounds from the second-heaviest child (Child #6), making this right-tailed difference in Weight greater than the entire Weight range for the other five children.

Child	Height (in)	Weight (lbs)
1	49	100
2	50	75
3	52	77
4	55	79
5	57	80
6	58	82

Table 12.1. Heights & Weights of Children – Weight-only outlier

In the table below, Child #1 is an outlier with respect to both Height and Weight. All children except for Child #1 in the table are close in Height differing at most by 8 inches and are close in Weight differing at most by 7 pounds. Child #1 is an outlier for both Height and Weight as the Height differs by 14 inches from the second-shortest child (Child#2) (which is greater than the entire Height range of the other five children), and the Weight differs by 18 pounds from the second-heaviest child (Child #6) (which is greater than the entire Weight range of the other five children).

Child	Height (in)	Weight (lbs)
1	36	100
2	50	75
3	52	77
4	55	79
5	57	80
6	58	82

Table 12.2. Heights & Weights of Children – Height and Weight outlier

Use the following stock price data (in dollars) for Exercises 13–18

10	**7**	**20**	**12**	**75**	**15**	**9**	**18**	**4**	**12**	**8**	**14**

Table A. Stock prices

13. Calculate the mean, median, and mode stock price.

The ***mean*** is calculated as the sum of the data points divided by the number of points as follows:

Mean Stock Price = (10+7+20+12+75+15+9+18+4+12+8+14) / 12 = 204 / 12 = $**17**.

The ***median*** is calculated by placing the prices in order and (a) selecting the middle value if the number of points is odd, or (b) taking the average of the two middle values if the number of points is even. Since we have twelve points, median is calculated as follows:

Median Stock Price = mean of center values {4,7,8,9,10,**12,12**,14,15,18,20,75} = 24/2 = $**12**.

The ***mode*** is calculated as the value that occurs the most often in the set and is calculated as follows:

Mode Stock Price = highest frequency of {4,7,8,9,10,**12,12**,14,15,18,20,75} = $**12**.

14. Compute the standard deviation of the stock price. Interpret what this number means.

The ***standard deviation*** represents the expected distance of a point chosen at random from a data set to the center of that set and is calculated by taking the square root of the ***variance***. The variance is the average of the sum of squared distances of each point from the data-set mean. Given that the mean is $17 (see Exercise #13) for this set, the variance for the set of stock prices is calculated as follows:

Stock Price Variance (Var) =

$(4-17)^2+(7-17)^2+(8-17)^2+(9-17)^2+(10-17)^2+(12-17)^2+(12-17)^2+(14-17)^2+(15-17)^2+(18-17)^2+(20-17)^2+(75-17)^2 =$

$(-13)^2 + (-10)^2 + (-9)^2 + (-8)^2 + (-7)^2 + (-5)^2 + (-5)^2 + (-3)^2 + (-2)^2 + (1)^2 + (3)^2 + (58)^2 =$

169 + 100 + 81 + 64 + 49 + 25 + 25 + 9 + 4 + 1 + 9 + 3364 = 3900 / 12 = **325 $^{2}**.

Taking the square root of the Variance, the Standard Deviation (SD) is calculated as follows:

Stock Price Standard Deviation (SD) of Stock Price = √(325) = **±$18.03**.

Since the mean is $17 and the standard deviation is plus/minus $18.03, the expected price of a stock drawn at random from the set of twelve stocks is expected to lie mathematically between ($17–$18.03) = **-$1.03** (i.e. $0.01 since we assume that a stock price can never be less than one penny USD) and ($17+$18.03) = $**35.03**.

As we can see, each stock with the exception of the one priced at $75 is priced within this range.

15. Find the min-max normalized stock price for the stock worth $20.

Min-Max normalization scales an observation relative to the data-set's range resulting in a value between 0 and 1 (this value has no units) and is formulated as follows:

$$\mathbf{MinMaxX_i = [X_i - Min(X)] / [Max(X) - Min(X)]}$$

Therefore, the min-max normalized stock price of $20 is calculated as follows:

MinMax($20) = ($20 - $4) / ($75 - $4) = ($16) / ($71) = **0.2254**.

Discovering Knowledge In Data: An Introduction to Data Mining, By Daniel T. Larose & Chantal D. Larose