Rich Gruss

Chapter 4 Important Points

Chapter Four delves into the process of cleaning data. Values are often missing or incorrect, and data is often too skewed to create meaningful analyses, so it helps to be armed with some strategies for transforming the data without altering it.

Some techniques of data transformation include discretizing continuous variables, normalizing and rescaling, and logarithmic transformations. *Discretizing* continuous variables is useful when the relationship between input and output isn't linear, but you're using a modeling technique that is. *Normalization* (dividing by median) is useful when absolute quantities are less meaningful than relative ones. The example in the book is age, because it's useful to compare an individual customer to the typical customer. *Log transformations* are helpful for data that is highly skewed. Money tends to be log normally distributed (i.e., log is normally distributed).

Sampling may not be necessary with today's computing power, but it still has its uses in testing and debugging because it's faster to work with a subset of data. Also, you can divide your data into training and testing splits. You train your model on the training set, and then validate it against the test (or hold-out) set to see how well it applies. R has a subset() function to make this easier.

Finally, Chapter 4 introduces and demonstrates how to use several slick functions, ,including is.na(), cut(), ifelse()—which we in the Java world called the ternary operator—merge(), and with(), which applies an expression against a data frame.