Corrections and Clarifications – Data Concepts and Visualization

p. 4.5: A WHERE statement is missing. After this line: FROM work_comp_claims, add the line: WHERE paid BETWEEN 1000 and 10000; (also adding semicolon after the statement)

p. 4.5 (middle of page - in paragraph following the statements): Replace sentence beginning "In SQL syntax…" with:

"In SQL syntax, single quote marks are used around the value, not the name of the data field; "back" is a value in the field, the field's name is injury."

p. 4.6: Please note that in some places the first letter of 'back' is not capitalized (such as 4.5) but in the table on 4.6, the first letter of 'Back' is capitalized. The reason this code will work is because many SQL implementations such as MySQL and Microsoft SQL server are case insensitive. This is in contrast to many languages such as R that are case sensitive to the values in character variables.

p. 4.12 (first line): A space is needed between SELECT and P and (on second line) between policy and P

p. 4.30 (in Knowledge to Action): A space is needed between ON and P on the first JOIN statement

p. 6.4 (in EDA and Metadata): On the last line of this section, change "number of records in the number of variables" to "number of records and number of variables."

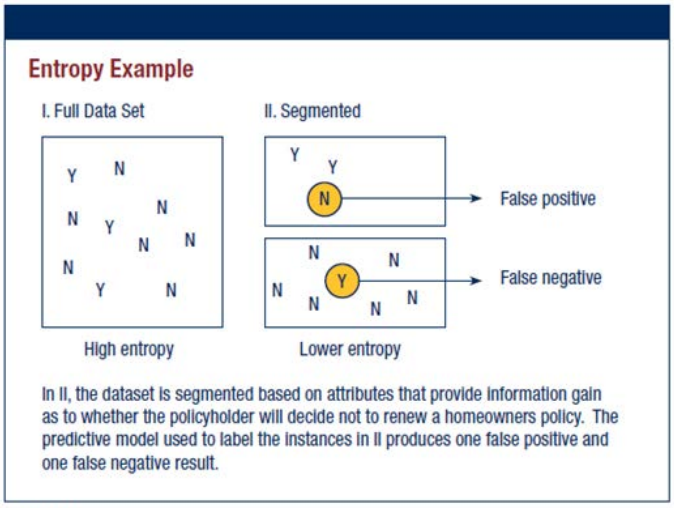
p. 6.10 (paragraph on Kurtosis): All instances should reflect that the actual kurtosis of the normal distribution is 3, not 4.

p. 6:32 (second heat map): Please note the circles should be proportionate in size to the percentages represented.

p. 6.52 (third from last paragraph from bottom of page): Change "how many exponents of 1" to "how many exponents of 10"

p. 7.5 (first line of formula at top of page): The term log (p_k) is missing the i subscript and so should be (p_{ik}).

p. 7.6: Delete the first full paragraph on this page (beginning with "Information gain…") and replace with this exhibit example (DA 11965) and these two paragraphs:



Information gain can be used to screen variables for a modeling data set. The "Entropy Example" exhibit is a visual representation of the entropy of the renewal/nonrenewal dataset. Each Y and N represents a class label of "yes, will renew," or "no, will not renew." When the dataset is segmented based on informative attributes (those with high information gain), the entropy decreases. To further clarify the example, see the data in the upcoming exhibit "Calculation of Chi-Squared Statistic Example." The entropy of the group with a length of time with the company of less than 5 years is 0.59 and for the group with more than 5 years is 0.91. The average entropy, when weighting each of these groups by their respective probabilities, is 0.69, which represents a 0.19 decrease in average entropy. The 0.19 represents the information gain.

Note that the information gain calculation was applied to categorical data. The "target" variable, renewal/non-renewal, is a binary categorical variable with two possible values, yes or no. In addition, the "predictor" variable, customer duration with the company, is also treated as categorical. That is, the numeric duration variable in the original data was split into two categorical groups. Then the entropy of data that has been classified based on the binary duration variable is calculated and used to compute and information gain.