Lecture and Assignment:

Clean Data!

Lecture Notes: What is Data cleaning, why is it important and how does one clean data?

Example: Have students open Clean Data set 1 in Canvas, then students produce an average of each column. (one erroneous bit of data can really change the answer. It would be good if student used Excel to complete the examples for practice.

Temperature (Fahrenheit)	Day length (hours)
70	9.8
78	9.9
98	10.0
95	10.1
76	10.1
99	10.1
079	10.2
88	10.3
79	10.5
140	6.01
Average ?	Average ?

Questions for review? What is the average of each column? Is this a "clean" average? Why or why not? If this is not representative of a clean average, what would make it more representative of the actual data?

Lecture notes: Data can be unclean or have errors if it has unwanted data, structural errors, non standardized data, unwanted outliers, or contradictory data errors. Give an example of each of these types of errors. Explain what each type represents and review why it changes the quality of data.

Why do we need to deal with these data cleaning issues in all data before we work with the information?

1. Impacts results
2. Rogue data may crash system
3. Garbage in Garbage out (GIGO) concept
4. Flaws in results or processes.
5. Keeps data organized
6. Avoids mistakes (customer to customer, field to field, product to product)
7. Improves productivity
8. Avoids unnecessary cost (overages, underage or incorrect product)

How do we deal with data cleaning basically?

Lecture notes: Look for errors visually, such as duplicate data, structural errors, (misspellings, incorrect capitalization, typos, rogue punctuation, incorrect label names) Structural errors are especially a problem in manually entered data. Look for standardized data errors, leading 0, inches v. centimeters, date inconsistency, et cetera.

Look for “outliers” and determine the best repair or cleaning. An outlier is a data point that is clearly outside of the norm (140 degrees Fahrenheit is an outlier in the example data set.)

Post the Averages Dirty Data Set 1 to Canvas, have students import and clean the data in Excell. Report what they did to clean the data, remove outliers, standardize data, inconsistencies,

Then create a function to merge the following two data sets.

Validate your data after the merger.

Data Cleaning tools, Microsoft Excel, (other spreadsheet software such as Google sheets) Programming languages (R, Ruby, SQL, Python), Graphic visualizations, (create a variety of graphs and look for outliers, in consistent labels, et cetera. Proprietary software (in this class we will use Ag. Leader and Excel generally.)

Discussion Question: What is the best response to missing data in a data set? Explain why you chose this answer and give examples.