Assignment 9

LPO 9951 / Fall 2015

The data you receive will rarely be perfect. Especially with such multi-dimensional collections of information, it is likely that systematic problems or anomalies will be present. There are a number of checks you will have to do and choices you will have to make to deal with the imperfections of large-scale datasets. Data cleaning is a skill that comes with lots of practice. Each dataset you encounter will have different quirks and a different level of "uncleanness."

For this assignment, narrow down your dataset to the variables that you will likely be using in your analysis. Write a one-page (single-spaced) summary of issues you identify that you will have to address in cleaning your data. Be sure to answer the following questions:

1. At first glance, are there any problems with the data? (Misspellings, incorrect data types, blatant anomalies?)

2. Do the logical relationships within the data make sense?

3. Do the basic descriptives for the data make logical sense?

4. Are there any patterns in the missing data?

5. How do you plan to deal with the problems you identify?

6. How might your approach affect the conclusions you can draw?

7. How have these issues been addressed in other research articles? (Identify an article that uses your dataset and note how the authors deal with data cleaning.)

8. What data issues are identified in the manuals and other documentation of your dataset?