Raw Data Checks  
07/10/2017

**GOAL: To understand raw data and its idiosyncrasies before implementing cleaning decisions.**

**Even before acquiring the data, you can ask:**

* What data are you expecting to receive?
  + From which sources? (SLDS, administrative, program, etc.)
  + In what format?
  + For which years?
* Is there any documentation (data dictionaries, codebooks, memos, etc.) that can help you understand the raw data, its structure, and its quality?
* Will you be able to link different tables together?
* Who can answer questions about the data?

**While exploring the raw data, you can ask:**

* What is the structure of each file? What is the level of uniqueness?
  + Are there duplicates or near-duplicates that need to be resolved?
* Does the number of records match what you expect? The number of unique students?
  + In total, by time period, by district or school, by subgroup, etc.
  + If the number is too low or inconsistent across years, is there a data quality or extract issue, or was there a true increase or decrease?
* What variables are there? Do you know what they are? Do you have all the variables you need?
* What are the patterns of missingness? Are there any variables that are entirely missing?
* Are the values of each variable consistent over time? Do you know what all the values mean?
* Do the rates and distributions of variables make sense?
* Can you link different tables together? How well do they merge?