cleaning framework

Interpretation of the data

- ▶What does each column / row represent?
- ▶What is the grain of the table?
- ▶Which columns are measures vs dimensions?

Clean and format the data

- ▶Format consistency
  - Dates MM/DD/YYYY
  - Numbers 77.77
- ▶Capitalization and abbreviations are consistent
- ▶Consolidate related fields
- ▶Check duplicate records

Identify missing and nonsensical data

- ▶Check nonsensical, blank, and NULL values
- ▶Which fields can be reasonably imputed?
- ▶Identify columns with missing data
  - Missing > 70%, likely unusable
  - Missing < 10%, likely usable
  - Use best judgement and make notes in write-up

Augment data

- ▶Add date fields for additional time grains
- ▶Add columns for join related fields using xlookup or vlookup
- ▶Calculate business metrics from one or more columns