Data Quality Checklist

Objective: Evaluate all data received for completeness, validity, consistency, accuracy, and interpretability in a timely manner so that:

- Basic and broad understanding of the data can be established as early as possible
- Problems with the data can be communicated to the client as early as possible, allowing sufficient flexibility in timeline should the data issues be severe enough to require re-extraction
- The client expectation can be managed effectively from the beginning

Key Things to Remember:

- Assume that all data are wrong until proven otherwise. Question everything.
- Do not try to interpret the data at this stage. This is a very fact-based exercise.
- No records are to be removed or contents altered with very few exceptions, until the final frequencies are confirmed with the client and the disposition of such data mutually agreed to. Absolutely no interpolation at this stage.
- Always consider the downstream impact of how you manage the data. Have justification for everything you do.

From Receipt[1]:	**At the Minimum[2]:**
First 24 Hours: Check *really* obvious facts	• Verify files and sources against data specifications in the project design document • Verify that the files can be uncompressed and opened without errors • Verify presence of data dictionary for each file where needed ○ Note that the layouts are absolutely required ONLY if the file has no column headers; however, the contents of the field may not be discernable from the column name alone • Anything else that are really obvious • Acknowledge receipt of the data and communicate any issues to client via email: ○ Confirm files received by name ○ List any missing files (check against the documented data requirements) ○ List any missing data dictionaries ○ State any issues with uncompressing and opening files ○ State any other really obvious issues ○ Otherwise state that a data quality analysis output will be delivered within 7 days barring critical data issues and will contact if any questions or issues come up in the meantime

[1] Of the respective file. These are milestones that should be reached at the end of the stated time, and it is expected that some work for the subsequent milestones is done in parallel as appropriate.

[2] Depending on the data or project, there may be other items that may need to be added to this list.

First 48 Hours: Check obvious facts

- Verify format of each file
- Create a ditto/xxd output (or equivalent) for each file
  - This is necessary especially for identifying non-displayable characters that impact your data load
- Visually inspect the data
  - Unless the files are small, create head and tail samples for each file
- Verify layout of each file: do not assume that the layout is correct
  - Delimited files: visually verify the field order against the raw data (not just the headers; also look at the contents), then verify that the length and the formats are correct
  - Fixed-column files: visually verify the positions and formats against raw data
- Verify record count of each file, using the line count or record count function that is native to the operating system
  - Do not use SAS/R to do this: Software tools have their own interpretation of what a “line” is, and must isolate facts from software interpretation
- Verify fields required against the documented data requirements
- Start to load raw files
  - As the files are loaded, check that the loaded record count matches the record count obtained above
  - Determine the best informat based on the layout, your verification of the layout, the messages SAS provides in the log, your understanding of the data, and common sense
  - Do not use PROC IMPORT
    - Assumes that file is clean, consistent, and relatively small—in practice this is rarely the case
  - Verify the width and the format of the field, even if they are given in a layout or data dictionary
    - They are not always correct
    - SAS could have its own interpretation which can vary from the definition native to the data source
    - For delimited records, use %varlen SAS macro/program
    - Considerations: character vs. numeric, number of decimal places, implied vs. explicit decimals, packed decimals, other non-displayable meta-characters, etc.
  - Additional things to consider:
    - If the same/similar fields are found in multiple files, should they all be read in with the same format/length?
    - How will these fields impact each other across different files?
    - What is the implication of reading each field in the format you have chosen?
    - Why do I need to load this line/file/variable/etc. this way?
- Note any issues in data load
- Check for presence of correct merge/unique key in each file
  - Based on the descriptions and the initial looks of the fields included, can tables be merged (theoretically)?
- Anything else that are obvious
- Communicate any issues to client via email

From Receipt[1]:	**At the Minimum[2]:**
First 72 Hours	• Complete the first round of data load • Check for exact duplicates o Do not remove them unless why they happen is known and it is confirmed appropriate to remove • Identify unique key(s) that uniquely identify a record o The purpose of the exercise is to understand the record level (i.e. what each record represents), as it often is NOT what you understand it to be o Start with your understanding of which fields/concepts would reasonably represent a record and their hierarchy o Use common sense. Just because a combination of fields uniquely identifies a record does not mean that the field is an appropriate ID variable. • Quantify the uniqueness of each ID (or ID-like) fields • Investigate any "duplicates" o NOTE: Nowhere does it say "remove duplicates" o Try to answer the question: what is different among the "duplicates"? • Check for "mergeability" o Do merge keys actually merge and how well do they merge? o This is further to the previous day where we simply answered whether it is theoretically possible) • Note any issues in data load • Check for general reasonableness of the sample received: o Volume o Target event rate or population average of the target metric, as applicable for the project (preliminary, or some other related metric) o Subpopulation count o Etc. This will vary by project. • Communicate any issues to client via email

From Receipt[1]:	At the Minimum[2]:
First 7 Days	• Finalize data load o At the minimum, your SAS log should be free of “error”, “warning”, “not initialized” (check both uppercase and lowercase)—none of these are acceptable. • Full frequency of every raw variable o Avoid shortcuts; do not assume a priori grouping is appropriate since it often hides problems. Instead, do a frequency on raw values with limited number of observations until you get a feel for the data, then apply groupings as necessary. o Work iteratively—this is a critical opportunity to see raw values of every variable with the naked eye. • An “eight-number summary” of every numeric variable for which this makes sense (#Missing, minimum, 1st percentile, 1st quartile, median, 3rd quartile, 99th quartile, maximum) o Note any potential default values o Consider whether any negative values make sense • Further check against the data requirements • Verify that the values in the distributions make sense • Verify that definitions are found for all coded values o Use the frequency above and label all coded values as they are found in the data dictionary and/or other relevant documents o Identify which values are missing descriptions, are inconsistent, etc. o Include all comments in the frequency output for client to evaluate (see below) • Verify that the distribution of key variables are reasonable • Verify consistency of fields within and across files (for example, similar fields from different data sources, class vs. subclass consistency, etc.) to the extent possible • Deliver a full report of the initial data check with comments and questions identified (an Excel file is generally acceptable especially at this stage): o File names with description and record count for each o Duplicate and merge check statistics o Frequency and distributions with variable definition and value descriptions indicated o All questions about the data o Comments on anything incomplete, unusual, unexpected, unknown, not understood, not explained, etc. o Comments on your initial insights about the data—validate your understanding of the data o Request an ETA for the answers to outstanding questions • Sign-off on the data if no uncorrectable error present; develop plan with timeline if uncorrectable error present or if the correctability is in question (e.g. requires client response to questions) o If the final sign-off on the file in question depends on the receipt of other files, then sign off conditionally and develop plan with timeline

From Receipt[1]:	**At the Minimum[2]:**
Questions to ask yourself prior to Day 7:	• Do I have the description for 100% of the fields? • Do I have all of the definitions? • Do I understand how the files relate to each other? • Do I understand everything in the data? Can I explain everything in the data? • Do the values make sense? • Are there any strange or unexpected values? • Are there any variables that should not be used from the practical perspective? • Am I ready to sign off on the data? (Once you sign off on the data, all fatal issues are your issues, not the client's!)
Questions to ask the client at Day 7:	• Any response to comments and observations (not just questions)? • Are there any variables that do not make sense? • Are there any variables that should not be used and why?