

## Case 19-8: Excel Data — Handout 1

### Excel Cleanup Guide

#### Benefits of Properly Prepared Data

- Data prepared in common and generally accepted format facilitate an expectation that a data file has a certain and predictable standard look, feel, and behavior.
- Excel analyzation tools work with properly prepared data in a predictable manner. For example, if a column contains a U.S. state and uppercase, lowercase, and proper case data are encountered, generating totals by state will reflect totals per case variations within the “state” column, and not by state.
- Certain Excel functions, such as XLOOKUP or VLOOKUP, may not perform properly if a cell is not formatted appropriately

#### Excel Data Preparation — Step-by-Step Instructions

##### *Step 1 — Establish Column Names (Fieldnames) in Row 1.*

Row 1 in Excel will contain the fieldnames used for performing audit procedures on digital data files. Use the following rules for fieldnames:

- The first character in the name must be alpha (A–Z).
- The rest of the fieldnames may contain 0–9, A–Z, or an underscore ( \_ ).
- No spaces or special characters like %, &, ©, \$, or # are allowed.

Abbreviate column names. Abbreviated means “as short as possible, yet recognizable.” Examples include DOCNO, CUSTNO, DESC, INV\_DATE, AMOUNT, ACCT, and PART\_NO.

##### *Step 2 — Make sure the first row of data appears in Row 2, immediately below the row containing fieldnames.*

##### *Step 3 — Eliminate nondata items, such as blank rows, columns, page breaks, headings or titles, subtotals, and totals.*

Below the column names in Row 1, the data should begin in Row 2 and be contiguous, with no blank rows or blank columns, subtotals or totals, or any nondata such as page breaks or report headings.

Consider sorting on a column with predictable content. For example: You receive a spreadsheet that looks like a report and contains page headers, page breaks, titles, column names, blank lines, dashed lines, totals, and subtotals repeatedly throughout. Column C of this spreadsheet is a customer number ranging from 000000 to 999999, and every row that contains an invoice has a customer number in Column C. Column C also contains “discard-lines,” which include items such as report headings, totals, subtotals, and blank lines; there are no 0–9 values. By sorting on Column C,

all the usable invoice data will fall into a large, contiguous block (000000–999999). Dashed lines will appear above that block, and all items such as blank lines, totals, subtotals, and headings, will fall below. Simply delete the dashed lines at the top, and all of the items such as headings, blank lines, totals, and subtotals, which fell below. Note that if the dataset contains empty rows, sorting alone may not be sufficient, as blank rows can interrupt the continuity of the data and result in fragmented blocks. To verify that the sorting works as intended, inspect and clean your data by removing or addressing empty rows before applying any sorting operations. The end result is column names in Row 1 and a contiguous block of data from Row 2 downward. Make sure the data still foot to your account balance being tested or other expected amount.

***Step 4 — Make sure dates and numeric data are formatted in a consistent manner. Date and numeric data format may vary across spreadsheets. The key is to ensure that data that will interact with one another are formatted consistently.***

*Numbers:* Format columns containing numbers using the 0.00 format. (Format cells > Number > Negative numbers: -1234.10.) Be sure to specify decimal places for consistent formatting.

*Dates:* Format dates using the date formatting convention of MM/DD/YY.

- Format cells > Date > Type: 03/14/12 > Locale (location): English (United States). For example, March 31, 2015, should look like 03/31/15. Note the leading zero.

***Step 5 — Make sure text columns are formatted in a consistent manner to ensure Excel functions execute in a consistent, predictable manner.***

Consider the following examples and how they would make data analytics using Excel more difficult.

- If you were to total the accounts receivable by state, and the state column contains uppercase, lowercase, and proper case data, you would get totals by each of the state variations that are based on case, not by state.
- If you were to perform an XLOOKUP or VLOOKUP in Excel, leading spaces (spaces on the left) would cause matching issues.
- Some data, such as account numbers and customer numbers, may physically be numeric in Excel but should be converted to text to perform key functions in Excel.

***Step 6 — Save your spreadsheet; you are ready for Excel functionality.***