Step-by-Step: Optimize the data-cleaning process

This reading outlines steps the instructor performs in the following video, Optimize the data-cleaning process. The video teaches some useful spreadsheet functions, which can make your data-cleaning even more successful.

Keep this step-by-step guide open as you watch the video. It can serve as a helpful reference if you need additional context or clarification while following the video steps. This is not a graded activity, but you can complete these steps to practice the skills demonstrated in the video.

What you'll need

If you would like to access the spreadsheet the instructor uses in this video, click the link to the dataset to create a copy. If you don't have a Google account, you may download the data directly from the attachments below.

Link to logistics data: International Logistics Association Memberships - Data for Cleaning

Link to cosmetics data: Cosmetics Inc. - Data for Cleaning

OR

International Logistics Association Memberships - Data for Cleaning
XLSX File

Cosmetics Inc. - Data for Cleaning
XLSX File

Example 1: The `COUNTIF` function

`COUNTIF` is a spreadsheet function that returns the number of cells within a range that match a specified value.

Use `COUNTIF` to find numbers lower than 100

1. Open the International Logistics Association Memberships - Data for Cleaning dataset, and scroll down to row 74.
1. **Note:** The dataset has 72 rows, and row 73 is left blank for separation.
2. In cell **H74**, enter **Member Dues < 100** to label the calculation.
3. In cell **I74**, enter the formula `=COUNTIF(I2:I72,"<100")` to count how many members in the cell range **I2:I72** pay dues of less than $100. This formula returns a value of 1, indicating one value is below $100.
4. In cell **I55**, change -$200 to $200. Cell **I74** should now display the value 0.

Use `COUNTIF` to find numbers higher than 500

1. In cell **H75**, enter **Member Dues > 500**.
2. In cell **I75**, enter the formula `=COUNTIF(I2:I72,">500")` to count how many members in cell range **I2:I72** pay dues of greater than 500. This formula returns a value of 1, indicating one value is above 500.
3. In cell **I44**, change $1,000 to $100. Cell **I75** should now display the value 0.

Example 2: The `LEN` function

The `LEN` function is useful if you have a certain piece of information in your spreadsheet that you know must contain a certain length.

1. Right click cell **A**.
2. Select + **Insert one column right** to create a new, empty column.
3. Select cell **B1** and enter **LEN** to name the new column.
4. In cell **B2**, enter `=LEN(A2)`. This function references the value of cell **A2** and returns its length, 6.
5. Double-click on the lower right corner of cell **B2**. This will copy the function through the rest of the column. Each cell will show the length of the Member ID in that row.

Example 3: Use conditional formatting

Conditional formatting is a spreadsheet tool that changes how cells appear when values meet specific conditions.

1. To highlight all of column **B** except for the header, select cell **B**. Then press CONTROL (Windows) or COMMAND (MAC) and select cell **B1**.
2. Navigate to the **Format** menu, and choose **Conditional Formatting**.
3. Set the **Format rules** field to **Is not equal to** and enter **6** as the value.
4. Select **Done**.
5. Notice cell **B36** is highlighted because its value is 7.

Example 4: The `LEFT` and `RIGHT` functions

`LEFT` is a function that returns a set number of characters from the left side of a text string. `RIGHT` is a function that returns a set number of characters from the right side of a text string.

The `LEFT` function

1. Use the Cosmetics Inc. - Data for Cleaning dataset.
2. Select cell **H1**, and enter **Left**.
3. In cell **H2**, enter `=LEFT(A2, 5)` to extract the first five characters from cell **A2**. This function will show the substring 51993.
4. Select cell **H2**.
5. Select and hold the fill handle, the small circle in the corner of a selected cell, then drag this formula down to populate the rest of this column.

The `RIGHT` function

1. Select cell **I1**, and enter **Right**.
2. In cell **I2**, enter `=RIGHT(A2, 4)` to extract the last four characters from cell **A2**. This function will show the substring Masc.
3. Select cell **I2**.

4. Select and hold the fill handle and drag this formula down to populate the rest of this column.

Example 5: The `MID` function

MID is a function that returns a segment from the middle of a text string.

1. Select cell **J1**, and enter **Mid**.
2. In cell **J2**, enter `=MID(D2, 4, 2)` to extract the two-letter state code that starts at character four in cell **D2**.
3. Double-click the fill handle and to automatically populate the rest of this column.

Example 6: The `CONCATENATE` function

`CONCATENATE` is a spreadsheet function that joins together two or more text strings.

1. Select cell **K1**, and enter **Concatenate**.
2. In cell **K2**, enter `=CONCATENATE(H2, I2)` to combine the values from columns H and I.
3. Double-click the fill handle and to automatically populate the rest of this column.

Example 7: `TRIM function`

`TRIM` is a function that removes leading, trailing, and repeated spaces in data.

1. Select cell **L1**, and enter **Trim**.
2. In cell **L2**, enter `=TRIM(C2)` to remove any leading, trailing, or repeated spaces.
3. Double-click the fill handle and to automatically populate the rest of this column.