# EXCEL

## quick data-cleaning tips

Created by Miranda Lee
January 2020

*This resource provides strategies for cleaning data in Microsoft Excel. Below is a brief overview of five situations you may find yourself in ("What") and corresponding solutions ("How"), followed by detailed instructions to implement the solutions.*

| | What? | How? |
|---|---|---|
| 1 | Identify all cells that contain a specific word or (short) phrase in a column with open-ended text | Use **Conditional Formatting** |
| 2 | Identify and remove duplicate data | Use **Remove Duplicates** function or **Conditional Formatting** |
| 3 | Identify the outliers within a data set (e.g., dates or grades) | Use **Data Validation** function |
| 4 | Separate data from a single column into two or more columns (e.g., first and last names) | Use **Flash Fill** |
| 5 | Categorize data in a column, such as class assignments or subject groups | Use **Formula** to fill in the category column |

---

### 1

**What: Identify all cells that contain a specific word or (short) phrase in a column with open-ended text**

**How:** *Use **Conditional Formatting** to locate cells that contain the desired word or phrase*

1. Highlight the column that has the data you want to search.
2. In the Home tab, click on **Conditional Formatting**, and then click **New Rule.**
3. Select **Format Only Cells That Contain** (on a Mac you need to select **Classic** from the first drop-down menu to see this option). Select **Specific Text**, **Containing**, and then enter the text you want to find. Adjust formatting options. Then click **OK.**
4. Any cells that contain the word or phrase you entered will now be highlighted.

Text "Data" ▶

| I love excel for data |
|---|
| Data cleaning is great |
| This is fun |

Highlighted "Data" ▶

| I love excel for data |
|---|
| Data cleaning is great |
| This is fun |

## 2 What: Identify and remove duplicate data

### How (Option 1): *Use the **Remove Duplicates** function*

1. Highlight the column or columns of data that you want to check for duplicate information.
2. In the Data tab, click **Remove Duplicates.**
3. If you had only highlighted one column, then just click **OK** in the Remove Duplicates window. If you highlighted more than one column, then you will have the option to select which columns to search for duplicates.
4. If you select more than one column in the Remove Duplicates window, then the data in all columns you select must be the same in order to be considered a duplicate.

Duplicates ▶

| | |
|---|---|
| Joe | Smith |
| Joe | Smith |

Not Duplicates ▶

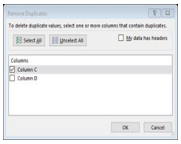
| | |
|---|---|
| Joe | Smith |
| Joe | Jones |

**Warning**: If you highlighted multiple columns of data, but then select only one column in the Remove Duplicates window, then Excel will do the following:

Original ▶

| | |
|---|---|
| Joe | Smith |
| Joe | Jones |

▶



Result ▶

| | |
|---|---|
| Joe | Smith |
| | |

### How (Option 2): *Use **Conditional Formatting***

1. Highlight the column of data you want to check for duplicates.
2. In the Home tab at the top of the Excel window, click **Conditional Formatting**, and then click **New Rule.**
3. Select **Format Only Unique or Duplicate Values** from the new drop-down menu (on a Mac you need to select **Classic** from the first drop-down menu to see this option). Then select **Duplicate**. Adjust format settings, then click **OK** once you have set the format.
4. Any cells that contain duplicate information will now be formatted according to the settings you selected. If you remove one item of a pair of duplicates, the relevant cells will revert to their original formats.
5. If you select multiple columns to find duplicates using this technique, Excel will consider each column separately.

Text "Joe" ▶

| |
|---|
| Joe Smith |
| Joel Smith |
| Dottie Smith |
| Joe Smith |

Highlighted "Joe" ▶

| |
|---|
| Joe Smith |
| Joel Smith |
| Dottie Smith |
| Joe Smith |

## 3 What: Identify the outliers within a data set (e.g., dates or grades)

**How:** *Use the **Data Validation** function to locate invalid data*

1. Highlight the column with the data you want to validate.
2. In the Data tab, click **Data Validation.**
3. In the Data Validation window, in the drop-down menu under "Allow," select the type of data you want to validate (e.g., whole numbers, dates, text length).
4. In the drop-down menu under "Data," select the appropriate condition (e.g., between, greater than, not between).
5. Enter the minimum and maximum valid data values in the appropriate boxes, then click **OK.**
6. Click the down arrow under the Data Validation button, in the **Data Tab**, and then click on **Circle Invalid Data**. Any data that does not meet the criteria you specified will then be circled.

**Cool tip**: Data validation can also help keep data sets clean. If you apply the Data Validation logic to empty cells, then you will get an error message if you try to enter invalid data into those cells.

2019 Dates ▶

| |
|---|
| 1/10/2019 |
| 3/21/2018 |
| 4/3/2019 |
| 11/26/2019 |
| 5/6/2018 |

Outliers Circled ▶

| |
|---|
| 1/10/2019 |
| 3/21/2018 |
| 4/3/2019 |
| 11/26/2019 |
| 5/6/2018 |

## 4 What: Separate data from a single column into two or more columns (e.g., first and last names)

**How:** *Use the **Flash Fill** function*

1. Insert a new column to the right of the column with the combined data.
2. Select the first cell in the new column and type the text from the combined column that you want to appear in your new column (e.g., the first name) and hit enter.
3. Repeat step 2 for a few cells. You should see the data autofill into the rest of the column based on the pattern you use. When this happens, just hit enter to finish the flash fill.
4. Add another column to the right of the new column for each piece of additional data you want to separate (e.g., the last name) and repeat this process.

**Cool tip:** You can use Flash Fill to separate any type of data that has a regular pattern (e.g., extracting the domain name from an email address). Create a new column, and then type everything before the @ when you do step 2 above.

It also works in reverse. If you have two columns next to each other, you can combine them by creating a new column and typing the information from each column in step 2 above.

One Column ▶

| |
|---|
| Dottie Smith |
| Harold Vicker |
| Gloria Perez |

Two Columns ▶

| | |
|---|---|
| Dottie | Smith |
| Harold | Vicker |
| Gloria | Perez |

5

**What: Categorize data in a column (e.g., class assignments or subject groups)**

**How:** *Use a **Formula** to fill in the category column*

1. Sort the raw responses so that similar responses are grouped together ("Data Column").
2. Add a column to the right or left of your data – this will be your "Category Column."
3. Place the appropriate category label next to the first row of each response type (i.e., Agree= positive, Disagree= negative, Strongly Agree= positive, Strongly Disagree= negative).
4. Highlight the cells in the Category Column, down to the bottom of the Data Column, and click the down arrow under **Find & Select** on the Home tab, and then click **Go to Special.**
5. In the Go to Special window, select the button next to **Blanks** and click **OK.**
6. You will see that the blank spaces are highlighted in gray, and the top empty space in the column is highlighted with a green border.
7. Without clicking anywhere, type an equal sign (=), and then select the cell above the green bordered column (should have the first category label in it).
8. Press the **Control** and **Enter** keys (on Mac, **Command and Enter**) at the same time to commit the formula to all the highlighted cells. The empty spaces should now be filled with the correct category labels.
9. Select the cells in the Category Column and press **Control** and **C** (on Mac, **Command** and **C**) at the same time.
10. Click the down arrow under **Paste** on the Home tab at the top of the Excel window.
11. Click the picture of the **Clipboard with the numbers 1, 2, and 3** on it (Paste values).

**Data**

| A |
|---|
| Agree |
| Agree |
| Strongly Disagree |
| Strongly Agree |
| Disagree |
| Agree |
| Agree |
| Disagree |
| Disagree |
| Strongly Disagree |
| Agree |

**Categories**

| |
|---|
| Positive |
| Negative |

**Groups**

| A | B |
|---|---|
| Agree | Positive |
| Agree | Positive |
| Agree | Positive |
| Agree | Positive |
| Agree | Positive |
| Disagree | Negative |
| Disagree | Negative |
| Disagree | Negative |
| Strongly Agree | Positive |
| Strongly Disagree | Negative |
| Strongly Disagree | Negative |

**Goal**

| A | B |
|---|---|
| Agree | Positive |
| Agree | Positive |
| Agree | Positive |
| Agree | Positive |
| Agree | Positive |
| Disagree | Negative |
| Disagree | Negative |
| Disagree | Negative |
| Strongly Agree | Positive |
| Strongly Disagree | Negative |
| Strongly Disagree | Negative |

EXCEL
quick data-cleaning tips

EvaluATE is supported by the National Science Foundation under grant number 1841783. Any opinions, findings, and conclusions or recommendations expressed on this site are those of the authors and do not necessarily reflect the views of the National Science Foundation.