


Tutorial: Introduction to Data Cleaning Using Excel



By the end of this lesson, you will be able to...

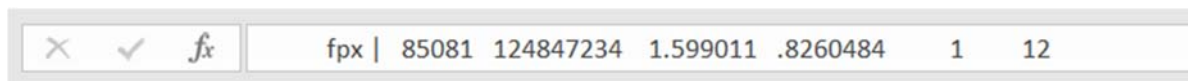
- Clean raw data using a variety of tools and shortcuts
- Create formulas and absolute references to other cells and worksheets
- Format Excel tables so that they are easy to read and useful for other users


Tip:

Any time you see a , be sure to pay careful attention. The icon signals a new skill or technique that will help facilitate your work!

Clean Raw Data

1. Open the “ExcelBasicsTutorial” workbook in Microsoft Excel. On the bottom of the screen, you should see six tabs (named “Final”, “Analysis”, “Raw”, etc). These “tabs” are called worksheets.
2. Navigate to the “Raw” worksheet. You can do this by directly clicking on the tab, or you can press  **Ctrl + PgDn** to toggle through the worksheets.
3. You should see some pretty messy output. The data you see is based on Stata code. I have pasted the code and the resulting tabulations in this worksheet. Our job now is to clean it up so that it's useful to us.
4. Let's start by converting the Stata output to table format. Click on cell **A9** (alternatively, you can hit  **F5** to bring up the **GoTo** window, type **A9** for the reference cell, and click **OK**). If you look in the formula bar at the top of the screen, you'll see that all of the data is in a single cell. As it is, there is no way to make this information useful to us.



5. But Excel has a useful tool to convert this nonsense into useful information. Make sure cell A9 is still highlighted. At the top of the screen, click on the **Data** ribbon. Click  **Text to Columns** under **Data Tools**.
6. On **Step 1** of the Wizard, select **Fixed width** and click **Next**. In the **Data Preview** window, click between columns to create a vertical separator between two columns (if the separators

are already in place, you can remove them by double-clicking on them, or move them around by dragging them). Once everything looks good, click **Finish**.

Convert Text to Columns Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

☐ Delimited - Characters such as commas or tabs separate each field.

☒ Fixed width - Fields are aligned in columns with spaces between each field.

Preview of selected data:

	ed_grp	Freq.	Percent	Cum.
200				
201				
202	Less than HS	1,536,838.5	20.78	20.78
203	HS Graduate or Equivalent	2,588,750	35.00	55.79
204	Some College	2,578,317.5	34.86	90.65
205	Bachelor's or Higher	691,560.5	9.35	100.00

Cancel < Back Next > Finish

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set field widths (column breaks).

Lines with arrows signify a column break.

To CREATE a break line, click at the desired position.

To DELETE a break line, double click on the line.


To MOVE a break line, click and drag it.

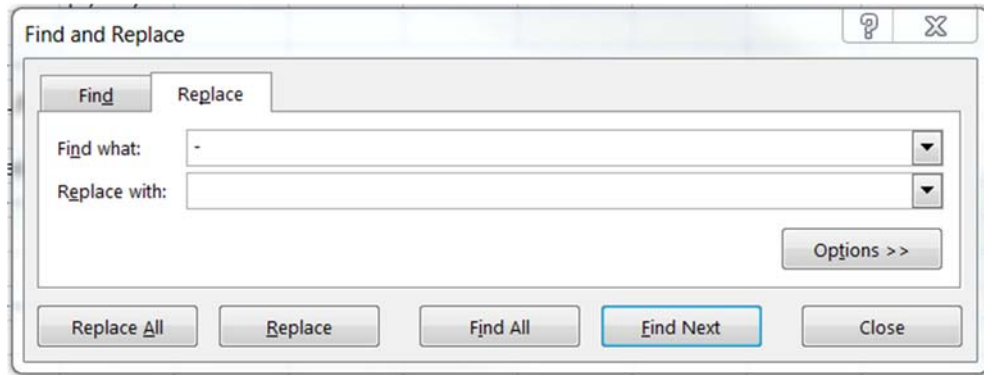
Data preview

	ed_grp	Freq.	Percent	Cum.
	Less than HS	1,536,838.5	20.78	20.78
	HS Graduate or Equivalent	2,588,750	35.00	55.79
	Some College	2,578,317.5	34.86	90.65
	Bachelor's or Higher	691,560.5	9.35	100.00

Cancel < Back Next > Finish


- You now have data separated into multiple rows and columns. To verify, navigate across the columns in row 9 using the arrow keys. Each value should now have its own cell.

8. But the data still looks a little messy, right? Let's get rid of all of the unnecessary symbols. Use  **Ctrl + H** to open up the **Replace** window. Here, type **+** into the **Find:** prompt. Leave the **Replace:** prompt empty. Now, click **Replace All**.
9. Repeat Step 8 using **- (minus)** and **| (pipe, or vertical bar)**, respectively.




10. If you'd like, you can repeat this step by highlighting rows 18 – 175 for only column A, and then again for only column K. If you don't need the extra practice, note that a fully cleaned version of the data can be found in worksheet "Raw(2)."
11. All right. Still messy, but now we can use it to create a table that will hopefully be useful.


Create a Table by Referencing Other Data

12. One benefit of creating a table in Excel over, say, Word, is that you can *reference* data in other cells to create data. This can get advanced pretty quickly, but we will deal with some basic examples here.
13. Navigate to the "Analysis" worksheet. Either using your mouse or **GoTo (F5)**, navigate to cell **C11**.  Here, type **=**. Now navigate to the "Raw" worksheet either using your mouse or the **Ctrl + PgDn** method [note: if you did not do Step 10 above, you should be using the "Raw(2)" worksheet instead of "Raw"]. Click on cell **C9**. Press **Enter**.
14. You will now be returned to the "Analysis" tab. You can now see the Estimate for the Total number of people age 25-54.
15. You should repeat this for each value in the table. I would suggest just doing a couple to get the hang of it and then move on to the next step.


Tip:

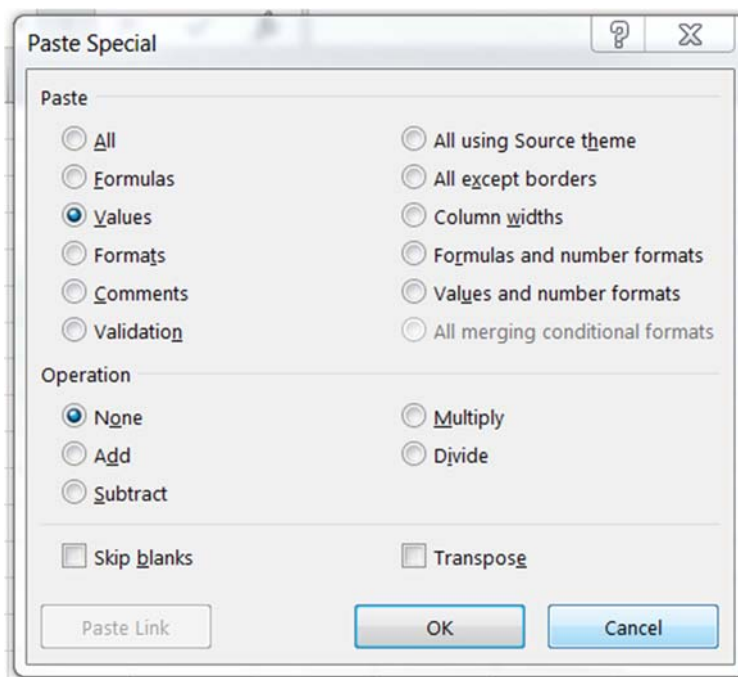
Not sure if you're referencing the right cells? Use  **Ctrl + ` (grave)** to display formulas instead of values.


You can compare your formulas to those in the “Analysis(2)” worksheet to double-check your work.

16. To calculate the “percent of total characteristic,” it might be easiest to use an example. Start in cell **D12**. Type “=”, and using the arrow keys navigate to cell **C12**. Then, type “/” and navigate to cell **C11** and press **Enter**. You should now have the percentage of whites as a percentage of all working age adults.
17. Right click on cell **D12** and select **Copy**. Then, right-click into **D13** and press paste. Alternatively, you can simply press **Ctrl + C** on **D12** and **Ctrl + V** on **D13**. This will copy the *formula* from the above cell. This way, you are now calculating the percent black instead of percent white.
18. Oh no! If you look at the formula, the denominator is the *Estimate of White, Non-Hispanic, not* the Total, which is what you want. Okay, this is how you fix it. Go to cell **D12**. In the formula bar, place the cursor anywhere in the denominator. Now,  click **F4**. You should now see a “\$” symbol before both the column and row number. This creates an *absolute* reference. It tells Excel that, no matter where you copy the formula, you want to use cell **C11** for the denominator.
19. Now, copy-and-paste-the formula all the way down the column. Instead of copying-and-pasting you can also drag the formula down. You can do this by placing your cursor on the bottom right-hand corner of **D12** until it forms a cross hair and then holding down the right-click on your mouse while dragging the cursor to the final row in the column. When you release, each cell should copy fill in the values based on the formula calculation.
20. Again, don’t feel pressured to fill in the entire table. Once you feel you have a good handle on the concept, skip to the next step.

Format Your Table

21. This is still a bit hard to read. So, we’ll go over a few tips to make the table more readable. First, use **Ctrl + A** to select all cells in the “Analysis” worksheet. Depending on where you clicked before selecting all, the command may have only selected the immediate table. If this is the case, just use **Ctrl + A** again do grab the entire sheet. Note: if you did not fill the entire table in Step 20, use the data from the “Analysis(2)” worksheet instead of “Analysis.”
22. Next, navigate to the “Final” worksheet. Since this is the sheet you will be presumably be sharing with colleagues, we do not want to simply paste everything. This will paste the formulas, which may just be confusing to other users. Instead, right-click on cell **A1**, navigate to  **Paste Special**. In the new window, select **Values** and click **OK**.



23. If you use **Ctrl + `**, you'll see that there are no formulas! Go back to the value view. Next, we're going to clean up the table a bit.
24. Let's start by clicking on the number 11 on the left-hand side. This selects the entire row with total values. Click on the "B" in the **Home** ribbon, or use **Ctrl + B** to bold the row.
25. Next, we're going to merge a few cells. Click on cell C9 and (still holding down on the mouse), drag the cursor to also select D9. Now, select  **Merge & Center** under **Alignment** in the **Home** ribbon. This merges both cells. Do the same for cells **E9** and **F9**.
26. You can also merge cells vertically. Try the same thing for cells **A12-A15**, **A16-A20**, and **A21-A24**.
27. Click cell **C9**. Under **Font in the Home** ribbon, click on the icon of a paint bucket (**Fill Color**) and select a color. Do the same for cell **E9**, but choose a different color.
28. Finally, select everything from **C9** to **D24**. Navigate to the box shape (**Outside Borders**) under **Font** in the **Home** ribbon. Select **Outside Borders**. Do the same for the following: **E9-F24**, **A12-F15**, **A16-F20**, **A21-F24**, and **A1-F24**.
29. And you're all done! Check the "Final(2)" worksheet to see how your table compares!

Raheem Chaudhry

Center on Budget and Policy Priorities
 rchaudhry@cbpp.org
 202.408.1080 x8352
 www.cbpp.org