# Hands-on Lab 5: Cleaning Data

Estimated time needed: 45 minutes

In this lab, first you will learn how to deal with inaccurate data, how to remove empty rows, and how to remove duplicated data. Next, you will learn how to change the case of text, how to change date formatting, and how to trim whitespace from data. Finally, you will learn how to use the Flash Fill feature and functions in Excel to help clean data.

# Software Used in this Lab

The instruction videos in this course use the full Excel Desktop version as this has all the available product features, but for the hands-on labs we will be using the free 'Excel for the web' version as this is available to everyone.

Although you can use the Excel Desktop software if you have access to this version, it is recommended that you use Excel for the web for the hands-on labs as the lab instructions specifically refer to this version, and there are some small differences in the interface and available features.

# Dataset Used in this Lab

The dataset used in this lab comes from the following source: [https://dataplatform.cloud.ibm.com/exchange/public/entry/view/f8ccaf607372882403a37d9019b3abf4](https://dataplatform.cloud.ibm.com/exchange/public/entry/view/f8ccaf607372882403a37d9019b3abf4?cm_mmc=Email_Newsletter-_-Developer_Ed%2BTech-_-WW_WW-_-SkillsNetwork-Courses-IBMDeveloperSkillsNetwork-DA0130EN-SkillsNetwork-20531059&cm_mmca1=000026UJ&cm_mmca2=10006555&cm_mmca3=M12345678&cvosrc=email.Newsletter.M12345678&cvo_campaign=000026UJ&cm_mmc=Email_Newsletter-_-Developer_Ed%2BTech-_-WW_WW-_-SkillsNetwork-Courses-IBMDeveloperSkillsNetwork-DA0130EN-SkillsNetwork-20531059&cm_mmca1=000026UJ&cm_mmca2=10006555&cm_mmca3=M12345678&cvosrc=email.Newsletter.M12345678&cvo_campaign=000026UJ). This dataset is published by IBM, and includes fictitious customer demographics and sales data.

We are using a modified subset of that dataset for the lab, so to follow the lab instructions successfully please use the dataset provided with the lab, rather than the dataset from the original source.

# Objectives

After completing this lab, you will be able to:

* Understand how to deal with irrelevant or inaccurate data
* Remove empty rows and duplicated data
* Change text case and date formatting
* Trim whitespaces from data
* Use Flash Fill and functions to clean data

# Exercise 1: Removing Duplicated, Irrelevant or Inaccurate Data

In this exercise, you will learn how to deal with inaccurate data, how to remove empty rows, and how to remove duplicated data.

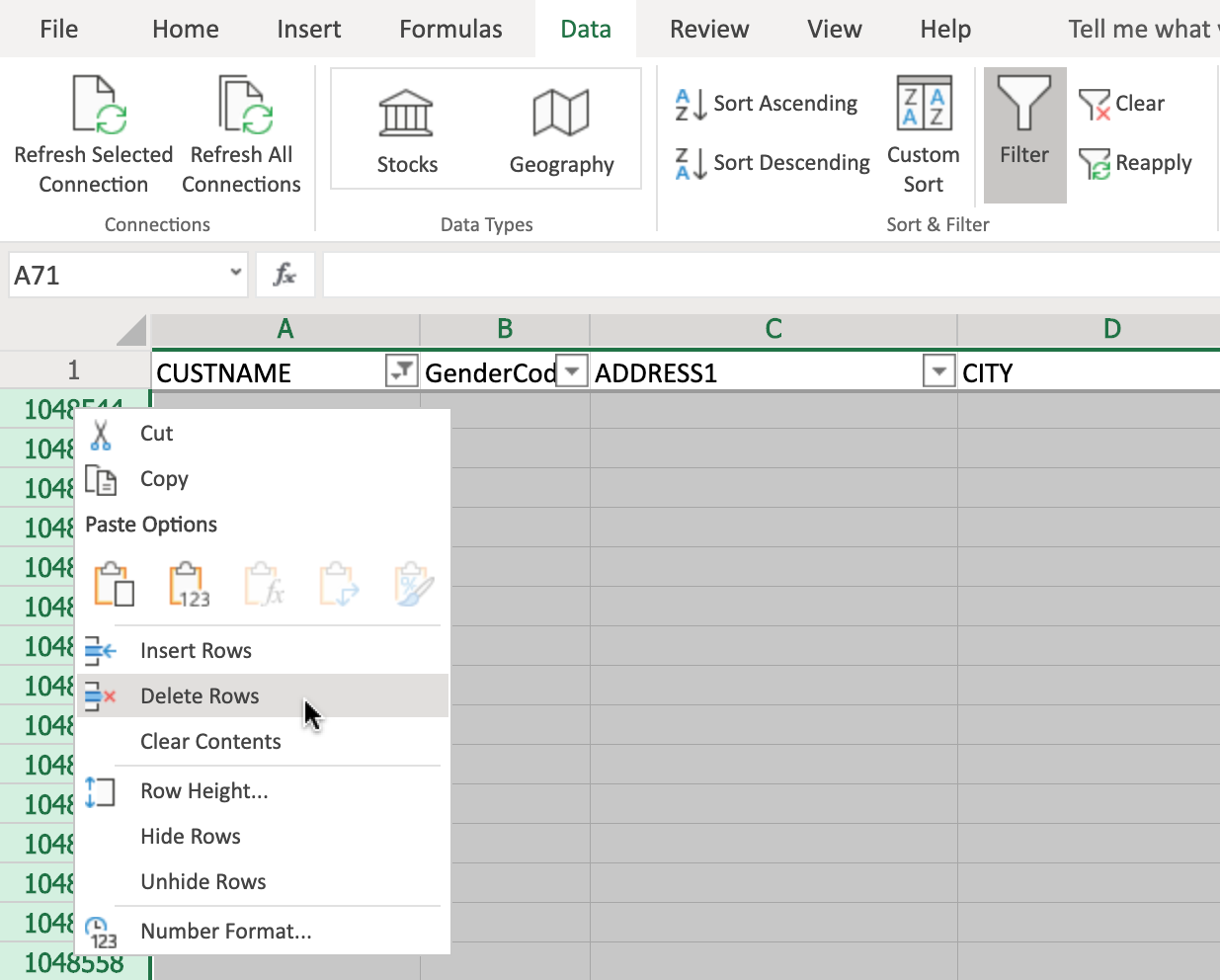
## Task A: Check spelling

1. Download the file [Customer\_demographics\_and\_sales\_Lab5.xlsx](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0130EN-SkillsNetwork/Hands-on%20Labs/Lab%205%20-%20Cleaning%20Data/Customer_demographics_and_sales_Lab5.xlsx). Upload and open it using Excel for the web.
2. Select column L (CREDITCARD\_TYPE), then click Review tab, and select Spelling.
3. Click the correct suggestion to change the spelling.
   * Note: Don’t change ‘jcb’ spelling when doing the spell check. We will need 'jcb' for the Exercise 1 Task D.
4. Close the Spelling pane.



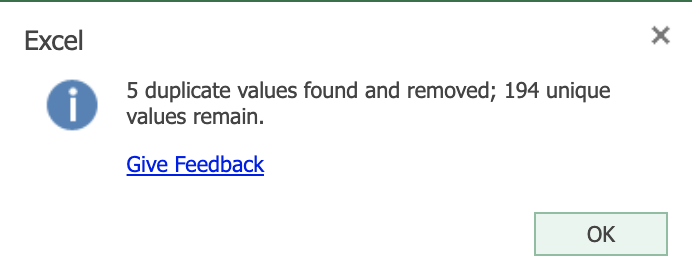
## Task B: Remove empty rows

1. Press CTRL+HOME, then press CTRL+SHIFT+END to select the whole datasheet.
2. On the Data tab, click Filter.
3. Press CTRL+HOME, click the filter arrow in the CUST\_NAME column, and then click Filter.
4. Click the Select All checkbox to deselect all of them. Then select just Blanks, then OK.
5. Select first row, then press CTRL+SHIFT+END to select all rows.
6. Right-click the selected rows and then click Delete Rows.
7. Finally, on the Data tab, click Clear, then click Filter.



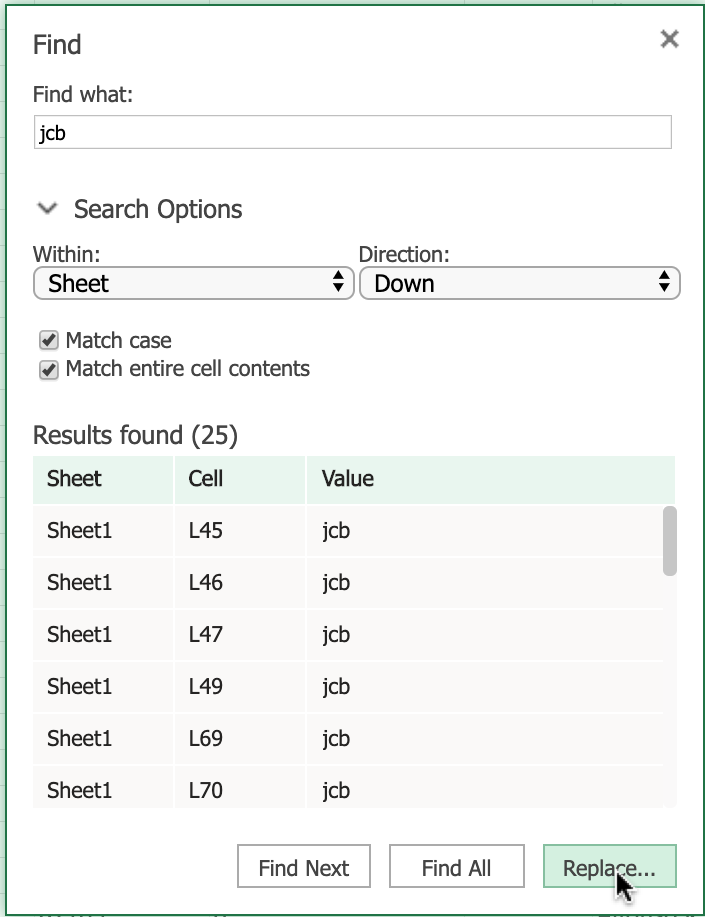
## Task C: Remove duplicate rows

1. Select Column T (ORDER\_ID) since ORDER\_ID values are unique.
2. On the Home tab, click Conditional Formatting> Highlight Cells Rules> Duplicate Values, and then click OK.
3. Select the whole datasheet (CTRL+SHIFT+END)
4. On the Data tab, click Remove Duplicates.
5. In the Remove Duplicates dialog box, ensure that Select all columns is checked and that My data has headers is also checked, then click OK.
6. In the pop-up box informing you how many duplicate values were found and removed, click OK.



## Task D: Use Find & Replace to correct misspelling

1. On the Home tab, click Find & Select.
2. Click Find. In Find what, type jcb, and click Find All.
3. Click Replace.
4. In Replace with, type JCB, click Replace All, and then click the Close icon.
5. On the Home tab, click Conditional Formatting> Clear Rules> Clear Rules from Entire Sheet.



# Exercise 2: Dealing with Inconsistencies in Data

In this exercise, you will learn how to change the case of text, how to change date formatting, and how to trim whitespace from data.

## Task A: Use the PROPER function to change text from upper case to proper case

1. Select row 2, then right-click it and choose Insert Rows.
2. In cell A2, type =PROPER(A1) and press Enter.
3. Hover over the bottom-right corner of cell A2, and drag the Fill Handle across to the last column.
   * If dragging across is too difficult with the mouse, then select the cells in the row 2 using SHIFT+RIGHT ARROW, then press F2 to put the cursor focus back in cell A2, then hold CTRL while you press Enter.
4. Select row 2, then press CTRL+C.
5. Select row 1, Right-click and choose Paste Options>Values.
6. Select row 2, right-click it and choose Delete Rows.

## Task B: Use the UPPER function to change text from proper case to upper case

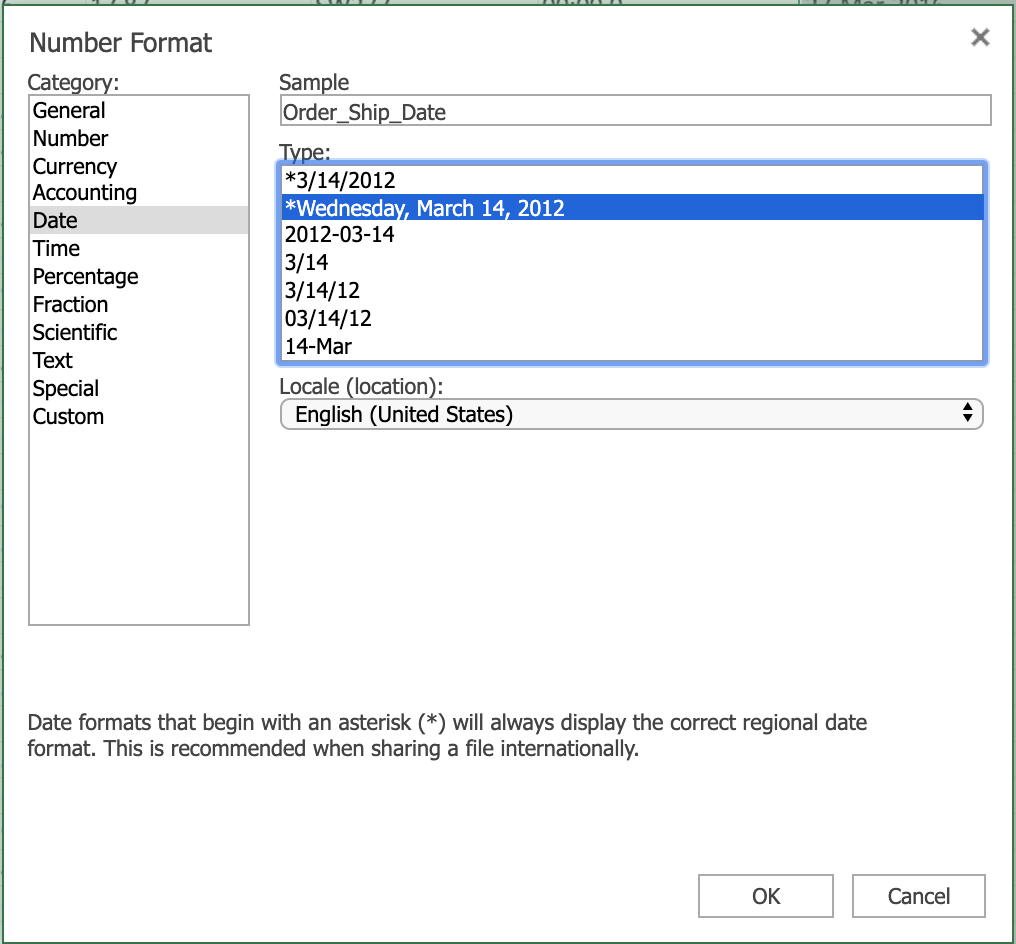
1. Select column AG (Generation). Then right-click and choose Insert Columns. In cell AG1, type Generation.
2. In cell AG2, type =UPPER(AH2) and press Enter.
3. Hover over the bottom-right corner of cell AG2 and double-click the Fill Handle.
4. Select column AG, then press CTRL+C.
5. Select column AH, right-click and choose Paste Options>Values.
6. Select column AG, right-click it and choose Delete Columns.

## Task C: Use the LOWER function to change text from proper case to lower case

1. Select column AC (T\_Type). Then right-click and choose Insert Columns. In cell AC1, type T\_Type.
2. In cell AC2, type =LOWER(AD2) and press Enter.
3. Hover over the bottom-right corner of cell AC2 and double-click the Fill Handle.
4. Select column AC, then press CTRL+C.
5. Select column AD, right-click and choose Paste Options>Values.
6. Select column AC, right-click it and choose Delete Columns.

## Task D: Change date formatting

1. Select column Z (Order\_Ship\_Date).
2. On the Home tab, in the Number group click Number Format> More Number Formats.
3. In the Category list, select Date.
4. In the Format Cells box, under Locale, select English (United States).
5. Under Type, select Wednesday, March 14, 2012 and click OK.



## Task E: Use Find & Replace to trim whitespace

1. Click CTRL+HOME.
2. Select all the data using CTRL+SHIFT+END.
3. On the Home tab, click Find & Select, then Replace.
4. In Find what, type 2 spaces. In Replace with, type 1 space.
5. Click Find All, then click Replace All.
6. Click the Close icon.

# Exercise 3: More Excel Features for Cleaning Data

In this exercise, you will learn how to use the Flash Fill feature and functions in Excel to help clean data.

## Task A: Use the Flash Fill feature to clean data:

1. Select column A (Cust\_Name), right-click and choose Insert Columns.
2. In cell A1 type Customer\_Name and press Enter.
3. In cell A2, type Mr. Allen Perl and press Enter.
4. Select column A (Customer\_Name), on the Data tab, click Flash Fill.
5. Click Undo to undo this step.

If you are using the desktop version of Excel, you could use the 'Text to Columns' feature to perform this next task (see the corresponding topic video for instructions).

If you are using ‘Excel for the web’ (the online version of Excel), the ‘Text to Columns’ feature is not available, but you can achieve the same results using functions, as shown in the steps below.

## Task B: Use LEFT, RIGHT, LEN, and SEARCH functions to clean data:

1. Select column A (Cust\_Name), right-click and choose Insert Columns.
2. Select column A again, right-click and choose Insert Columns.
3. In cell A1, type Customer\_Firstname and in cell B1, type Customer\_Lastname.
4. Click C1, then on the Home tab, click Format Painter, then drag across to A1 and B1.
5. Double-click the divider between columns A and B.
6. In cell A2 type =LEFT(C2, SEARCH(" ",C2,1)) and press Enter.
7. In cell B2 type =RIGHT(C2,LEN(C2)-SEARCH(" ",C2,1)) and press Enter.
8. Double-click the Fill Handle on cell A2.
9. Double-click the Fill Handle on cell B2.

### Congratulations! You have completed Lab 5, and you are ready for the next topic.