Unit 3

Data Cleaning

**What Is Data Cleaning in Excel?**

**Data Cleaning in Excel**is a combination of processes that includes identifying, highlighting, and acting against all kinds of errors to have organized data. There are many built-in Excel functions and features that are directly or indirectly used to ease the data-cleaning process in Excel.

## Why Is It Important to Clean Data?

As we use data for many important decisions, it is a must to keep the data clean as far as possible. Some of the reasons are listed below:

* **Accuracy:** Ensures better accuracy while data is analyzed.
* **Reliability:** Instinctively enhances reliability.
* **Efficiency:** Reduces time spent on it as all the data are already sorted and ready to use.
* **Better Decision-making:** Less error-prone and helps to predict better decisions.

## What Are the Common Issues with Data That Need Data Cleaning?

Data cleaning in Excel is a very vast section. Some of the most common issues with data that need cleaning include:

* Spelling mistakes
* Duplicate data
* Uncategorized text cases
* Unnecessary spaces, non-printable characters, and formats
* Unorganized formatting of numbers, times, and dates
* Missing text in a specific position, etc.

## What Are the Basics of Data Cleaning in Excel?

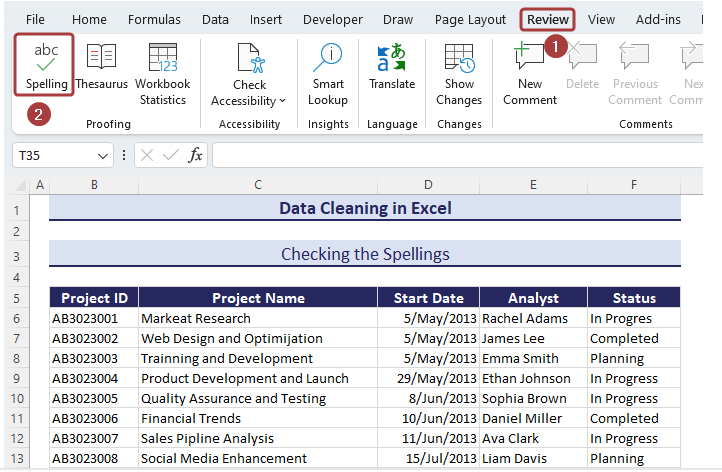
The basics of data cleaning in Excel can be summed to these few steps:

* Import the raw data from an external data source.
* Keep a copy of the imported data in a separate workbook.
* Organize the data in table form with the right data in the right place.
* Make the data cleaning according to your needs.

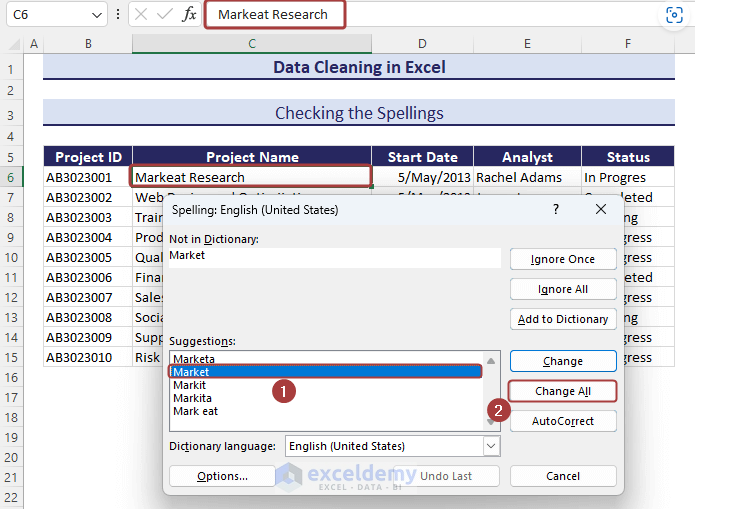
## How to Clean Data in Excel?

### Method 1 – Spell-Checking

* Go to the **Spelling**option from the **Review**tab.

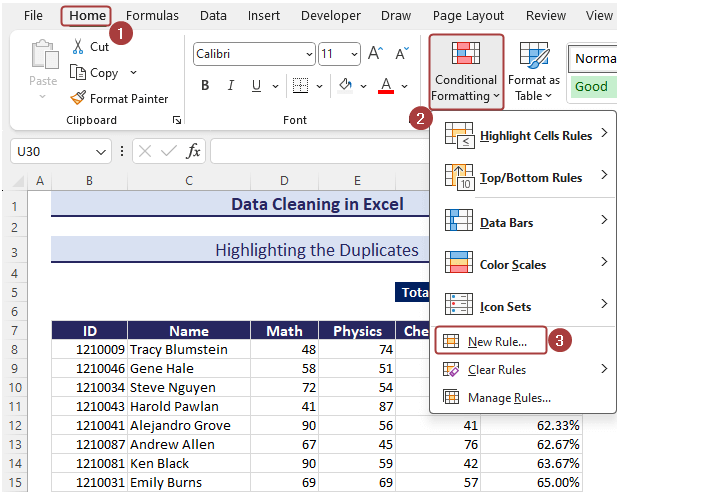


* A wizard named **Spelling: English (United States)** will appear with necessary suggestions of the misspelt words.
* Pick the correct spelling from the available spelling suggestions and click on **Change All**to make corrections in the entire worksheet.



### Method 2 – Highlighting Duplicates

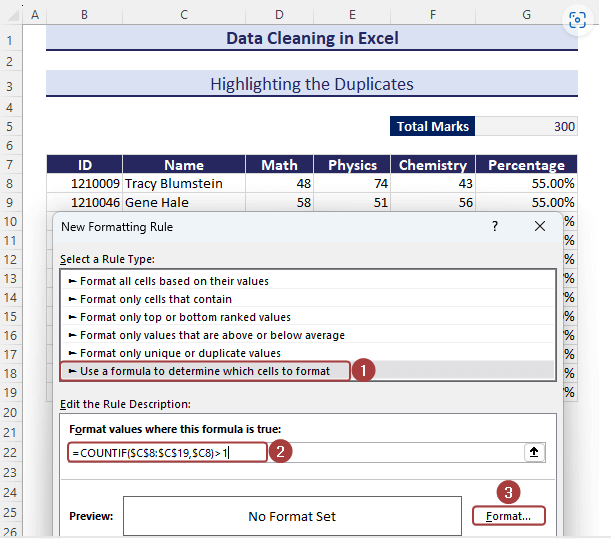
* Go to **Conditional Formatting**from the **Home**tab.
* Select the **New Rule…**option.



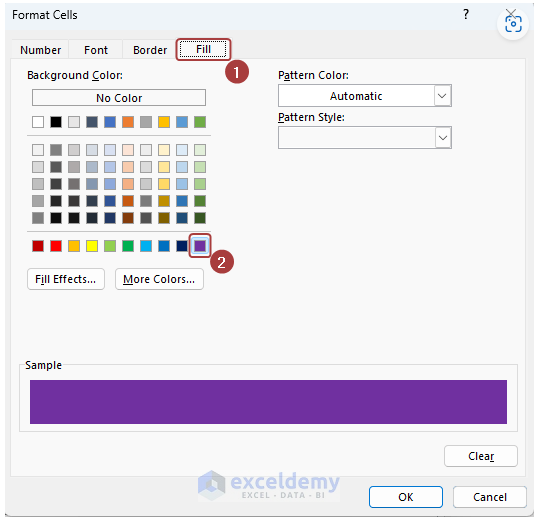
* Pick the **Use a formula to determine which cells to format**option from the **New Formatting Rule**wizard.
* Insert the following formula in the **Format values where this formula is true**section:

**=COUNTIF($C$8:$C$19,$C8)>1**

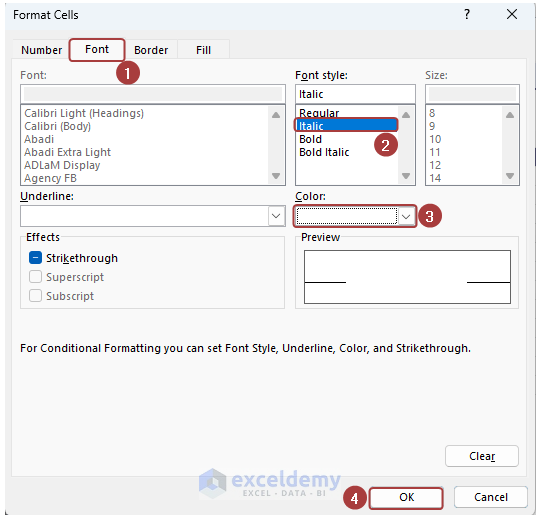
* Click on the **Format**option to define the matched values format.



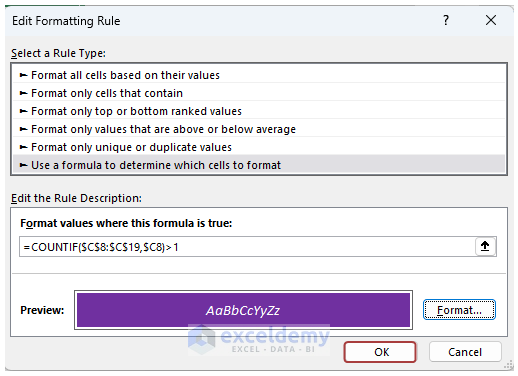
* Pick a background color from the **Fill**tab.



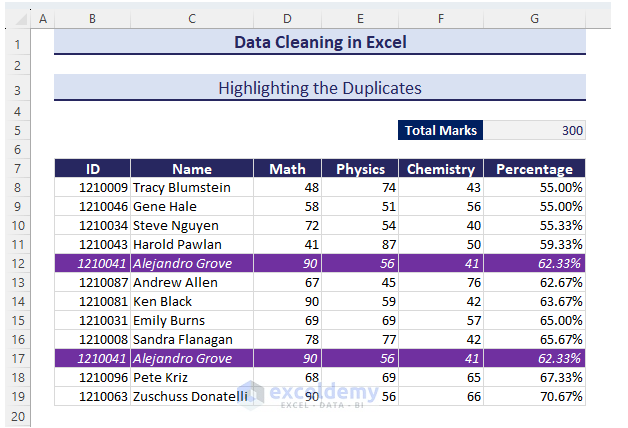
* You can customize the font style of the matched values. We have set the font color white and style italic for the matched cells.
* Click **OK**to finish the formatting.



* Click **OK**again to apply the conditional formatting.

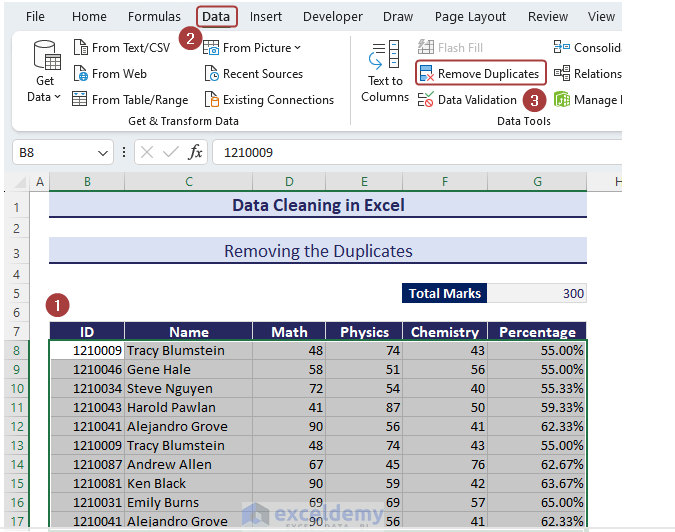


* We have the duplicate values highlighted according to the defined formatting.

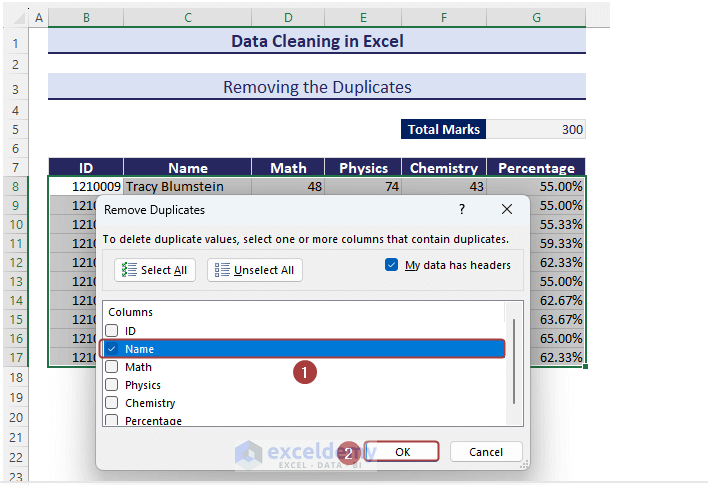


### Method 3 – Removing Duplicates

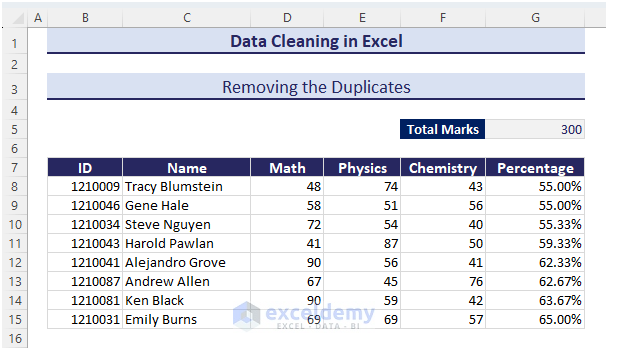
* Select the entire data and go to the **Data**tab.
* Click on **Remove Duplicates**.



* Pick a column and click on **OK**to find the duplicates and delete the entire row.

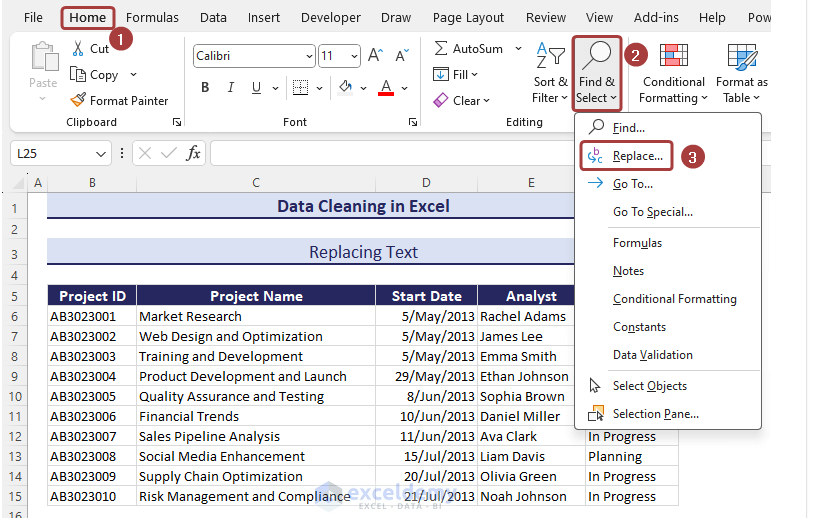


* We’ll get a dataset with no duplicates along the defined column.

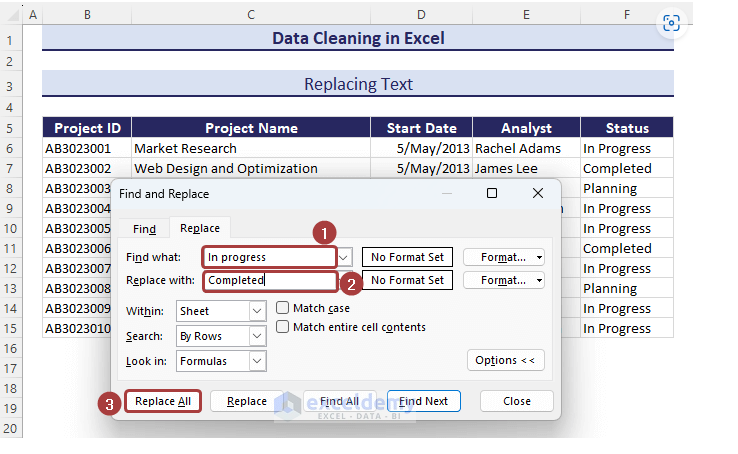


### Method 4 – Replacing Text

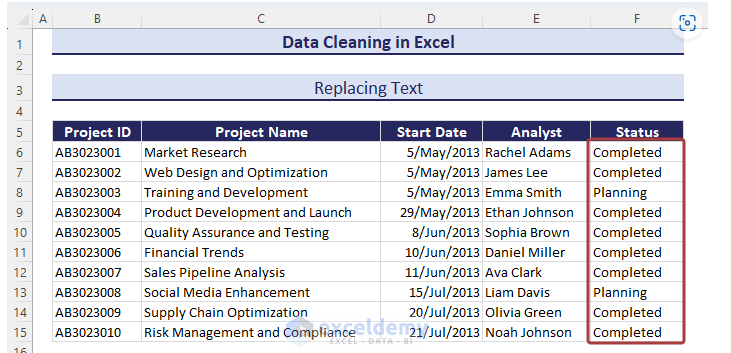
* Go to the **Find & Select**command from the **Home**tab.
* Pick **Replace…**from the available options.



* Input the text to be replaced (i.e. ***In Progress***) in the **Find what**section and the text that will be inserted (i.e. ***Completed***) in the **Replace with**section.
* Click on **Replace All**.



* Thus, we can simply replace texts.

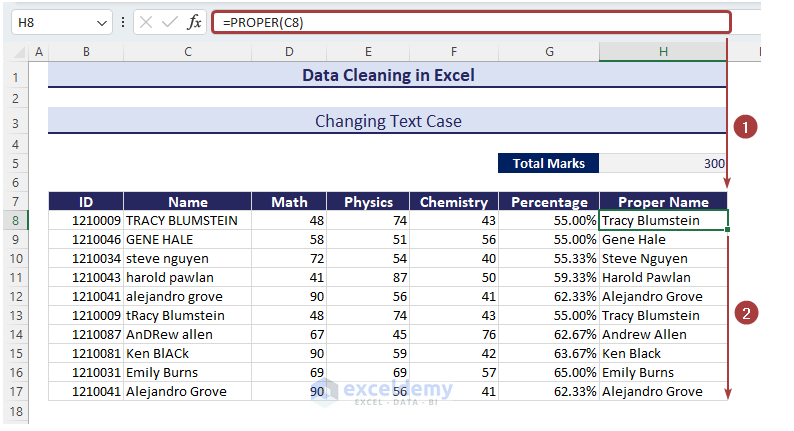


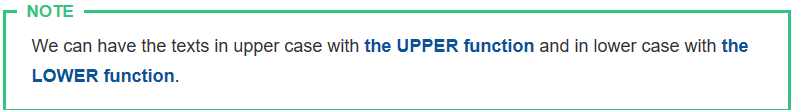
### Method 5 – Changing Text Cases

* Apply the following formula with [**the PROPER function**](https://www.exceldemy.com/excel-proper-function/) in cell **H8** to have the name of cell **C8**in the proper case:

**=PROPER(C8)**

* Use the **Fill Handle**to autofill the formula.

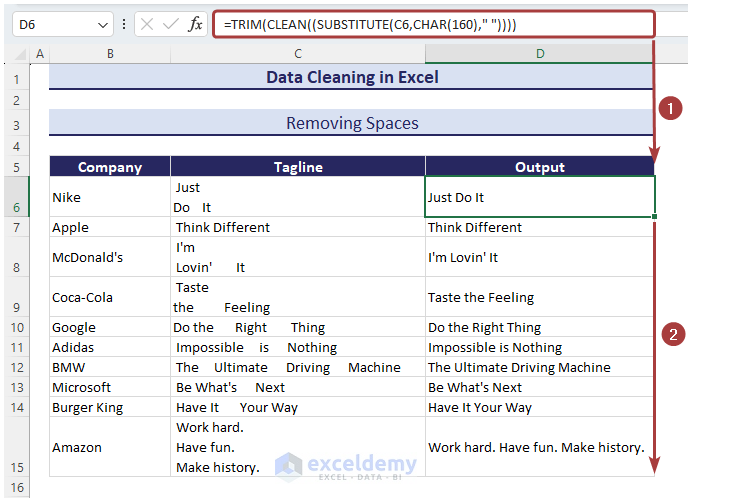




### Method 6 – Removing Spaces

* Apply the following formula with the [**TRIM**](https://www.exceldemy.com/excel-trim-function/), [**CLEAN**](https://www.exceldemy.com/excel-clean-function/), and [**SUBSTITUTE**](https://www.exceldemy.com/excel-substitute-function/)functions to remove spaces between texts as well as leading spaces at the beginning:

**=TRIM(CLEAN((SUBSTITUTE(C6,CHAR(160)," "))))**



### Method 7 – Removing Non-Printable Characters

Look at Column C, under the **Name**header, and you’ll find some unwanted characters before or between texts that are not printable.

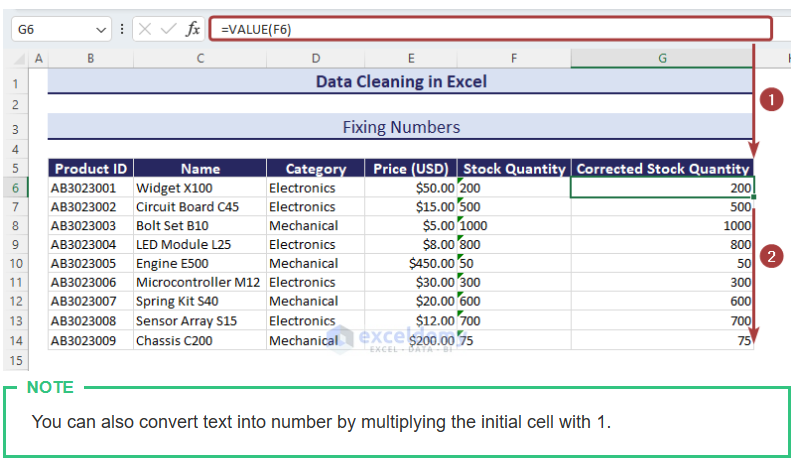
* Apply the following formula with [**the CLEAN function**](https://www.exceldemy.com/excel-clean-function/) to have the non-printable characters removed:

**=CLEAN(C8)**

### Method 8 – Fixing Numbers

* Use the following formula with [**the VALUE function**](https://www.exceldemy.com/excel-value-function/) to put the numbers in the number format.

**=VALUE(F6)**

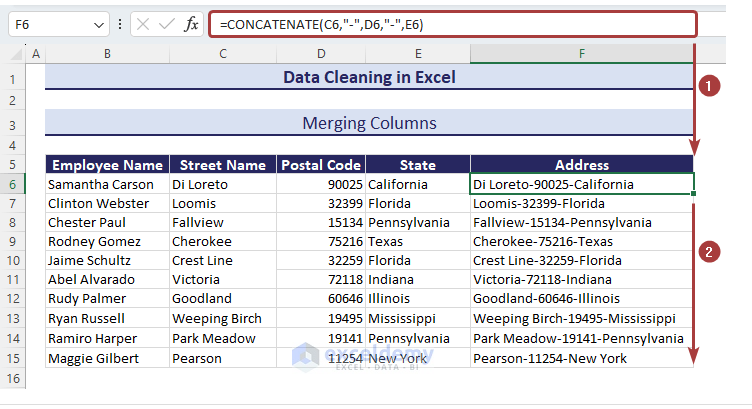


### Method 9 – Merging Columns

Under the **Address** header in column F, we’ll show the proper address format by merging the street name, postal code, and state name from the left 3 columns.

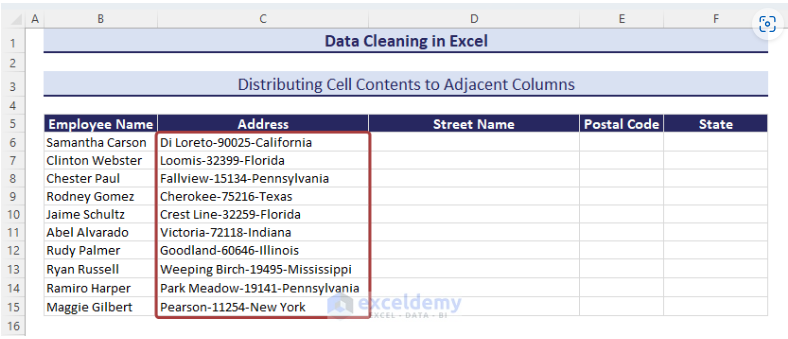
* Apply the following formula with [**the CONCATENATE function**](https://www.exceldemy.com/excel-concatenate-function/) to merge the columns and separate the segments with dashes:

**=CONCATENATE(C6,"-",D6,"-",E6)**

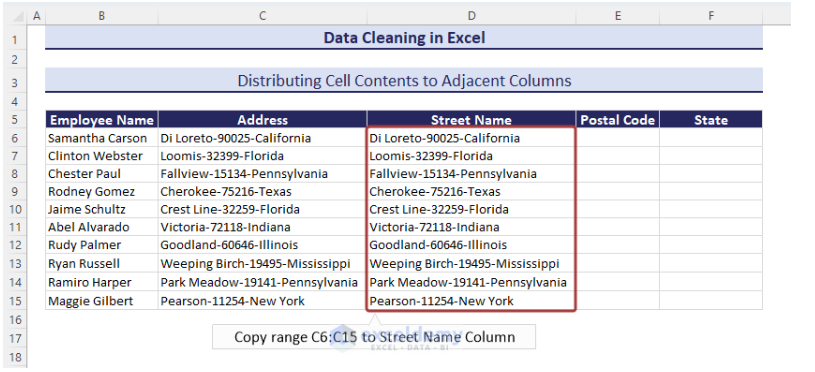


### Method 10 – Distributing Cell Contents to Adjacent Columns

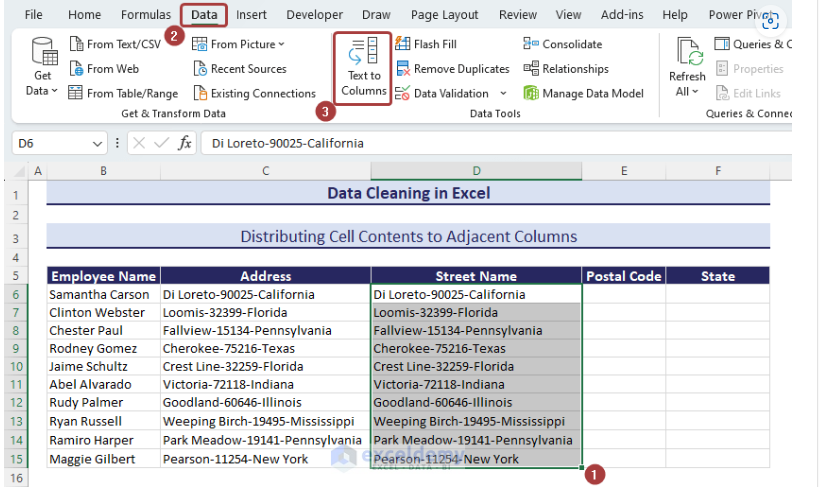
We have a dataset where the addresses are the combination of the street name, postal code number, and state name. Those are separated with a dash (**–**) between them.



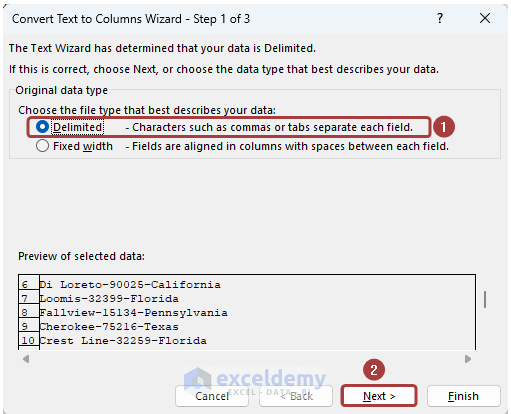
* Copy the entire column values to the **Street Name**column.



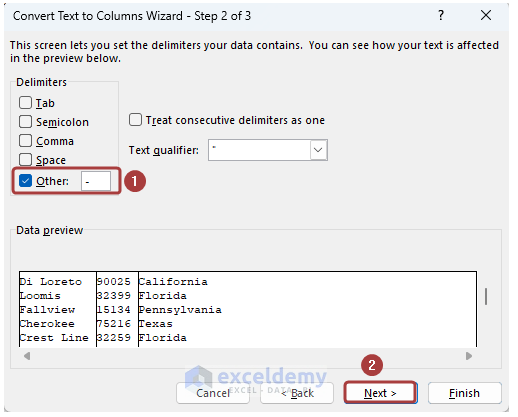
* Select all the values in the **Street Name**column and click on the **Text to Columns**option from the **Data**tab.



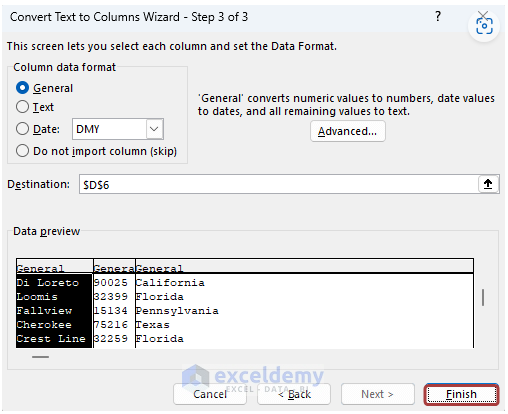
* From the **Convert Text to Columns Wizard**, choose the **Delimited**since the data is combined with the dash sign.
* Click on **Next**.



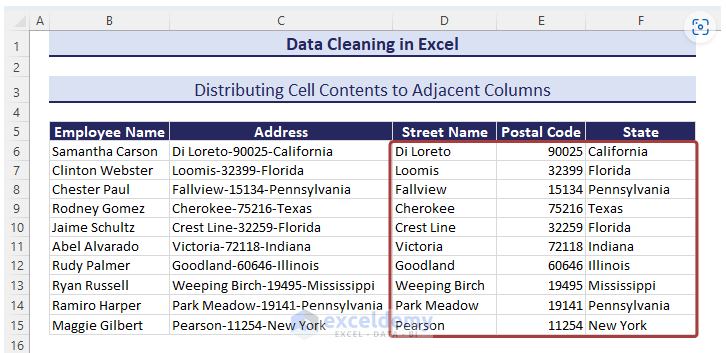
* Define the delimiter based on what the cell values are separated.
* Click on the **Next**button.



* Click on **Finish**to end the process.

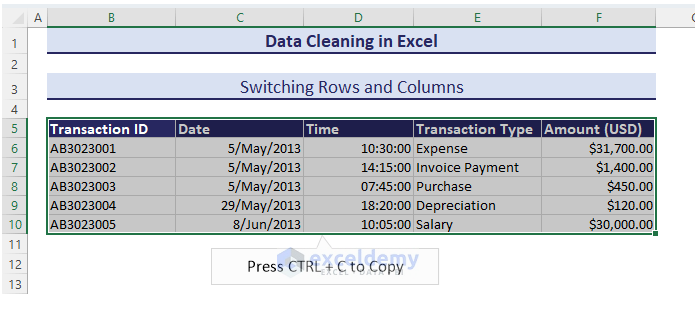


* We have the distributed cell values in the adjacent cells.

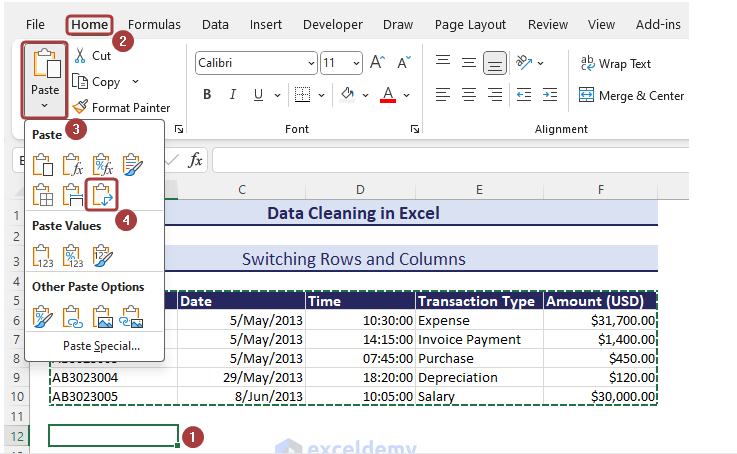


### Method 13 – Switching Rows and Columns

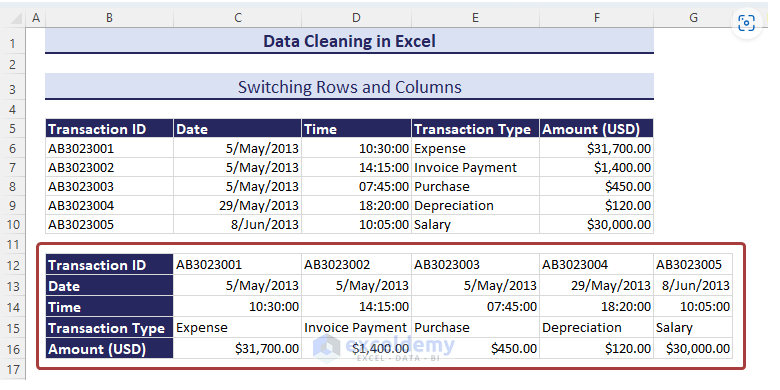
* Copy the entire range.



* Select a cell to paste the switched rows and columns.
* Go to **Paste**from the **Home**tab.
* Pick the **Transpose (T)**option to make the switch



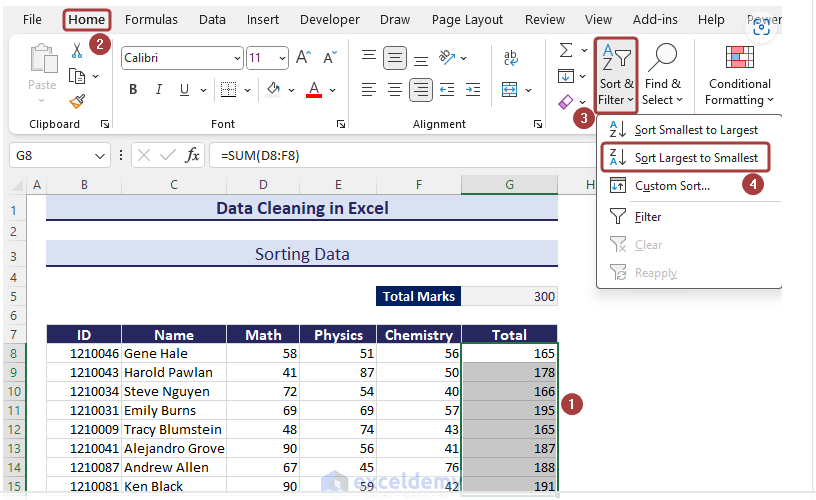
* We’ll get the switched rows and columns.



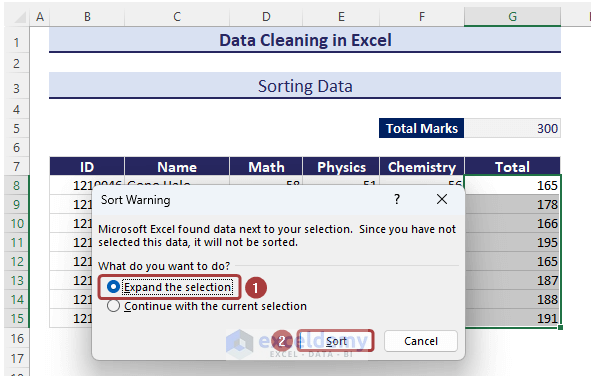
### Method 17 – Sorting Data

In the following dataset, we will sort data in descending order based on the total marks.

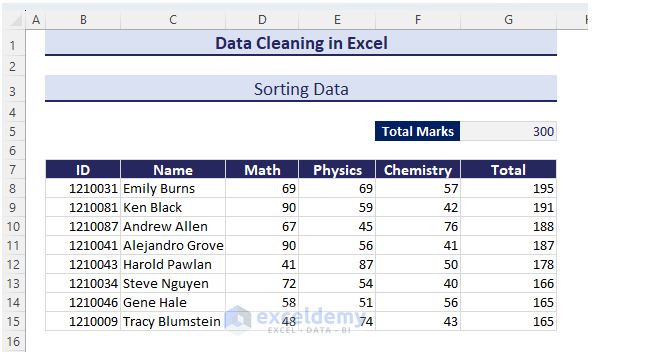
* Select the column you want to sort by.
* Go to the **Home**tab and click on **Sort Largest to Smallest**from **Sort & Filter**to sort in descending order.



* Choose the **Expand the selection**option and click on **Sort**.



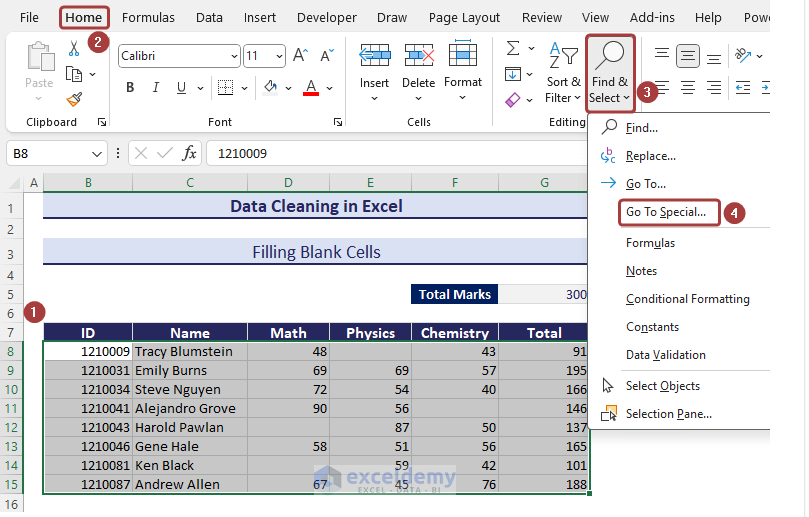
* We have the table sorted in descending order.



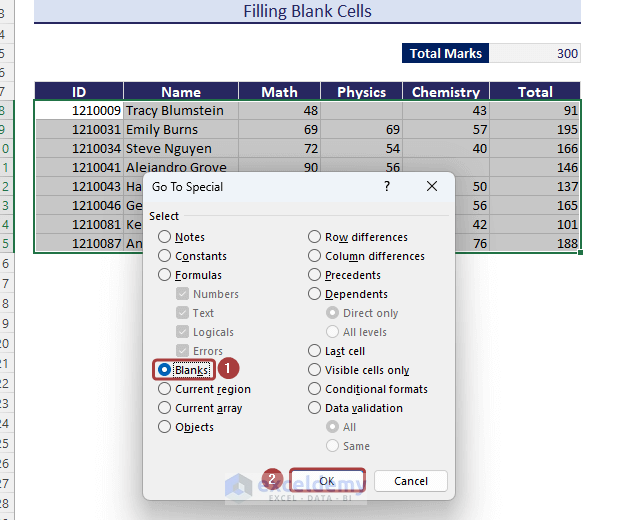
### Method 20 – Filling Blank Cells

Blank cells make a dataset unfulfilled. We can insert zeros in those cells to have a better representation.

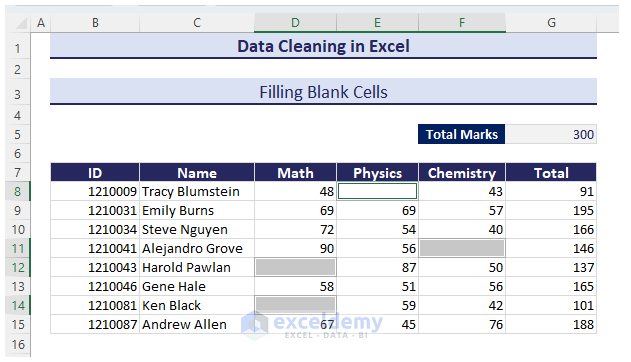
* Select the entire range.
* Go to the **Home**tab and select **Find & Select**from the ribbon.
* Pick **Go To Select…**from the available options.



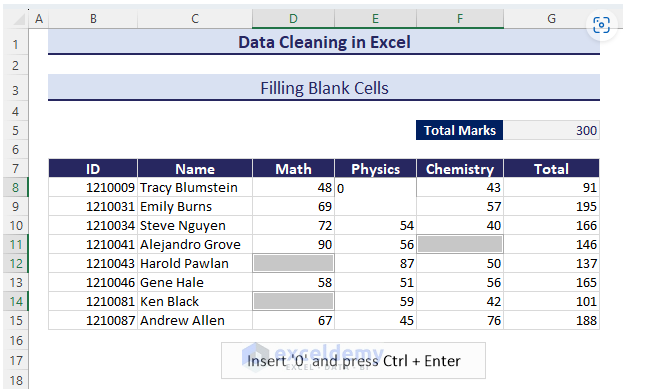
* Select **Blanks**and click on **OK**.



* The blank cells within the selected range will be selected.



* Insert zero and press **Ctrl + Enter**.



We will have the blank cells filled with zeros

