# Cleaning Data with Text Functions

The data that you obtain from different sources many not be in a form ready for analysis. In this chapter, you will understand how to prepare your data that is in the form of text for analysis.

Initially, you need to clean the data. Data cleaning includes removing unwanted characters from text. Next, you need to structure the data in the form you require for further analysis. You can do the same by −

- Finding required text patterns with the text functions.
- Extracting data values from text.
- Formatting data with text functions.
- Executing data operations with the text functions.

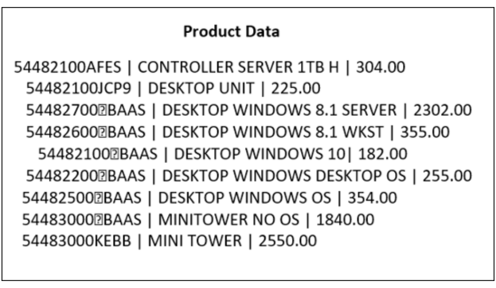
## Removing Unwanted Characters from Text

When you import data from another application, it can have nonprintable characters and/or excess spaces. The excess spaces can be −

- leading spaces, and/or
- extra spaces between words.

If you sort or analyze such data, you will get erroneous results.

Consider the following example −

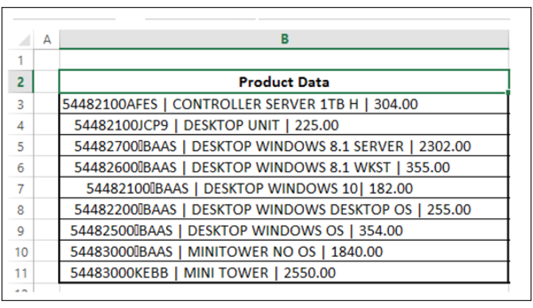


**Product Data**

54482100AFES | CONTROLLER SERVER 1TB H | 304.00
54482100JCP9 | DESKTOP UNIT | 225.00
54482700□BAAS | DESKTOP WINDOWS 8.1 SERVER | 2302.00
54482600□BAAS | DESKTOP WINDOWS 8.1 WKST | 355.00
54482100□BAAS | DESKTOP WINDOWS 10| 182.00
54482200□BAAS | DESKTOP WINDOWS DESKTOP OS | 255.00
54482500□BAAS | DESKTOP WINDOWS OS | 354.00
54483000□BAAS | MINITOWER NO OS | 1840.00
54483000KEBB | MINI TOWER | 2550.00

This is the raw data that you have obtained on product information containing the Product ID, Product description and the price. The character "|" separates the field in each row.

When you import this data into Excel worksheet, it looks as follows −



| | A | B |
|---|---|---|
| 1 | | |
| 2 | | **Product Data** |
| 3 | | 54482100AFES \| CONTROLLER SERVER 1TB H \| 304.00 |
| 4 | | 54482100JCP9 \| DESKTOP UNIT \| 225.00 |
| 5 | | 54482700□BAAS \| DESKTOP WINDOWS 8.1 SERVER \| 2302.00 |
| 6 | | 54482600□BAAS \| DESKTOP WINDOWS 8.1 WKST \| 355.00 |
| 7 | | 54482100□BAAS \| DESKTOP WINDOWS 10\| 182.00 |
| 8 | | 54482200□BAAS \| DESKTOP WINDOWS DESKTOP OS \| 255.00 |
| 9 | | 54482500□BAAS \| DESKTOP WINDOWS OS \| 354.00 |
| 10 | | 54483000□BAAS \| MINITOWER NO OS \| 1840.00 |
| 11 | | 54483000KEBB \| MINI TOWER \| 2550.00 |

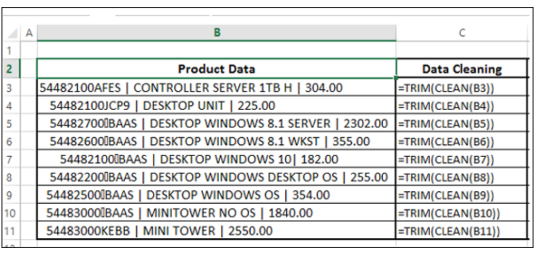
As you observe, the entire data is in a single column. You need to structure this data to perform data analysis. However, initially you need to clean the data.

You need to remove any nonprintable characters and excess spaces that might be present in the data. You can use the CLEAN function and TRIM function for this purpose.

| S.No. | Function & Description |
|---|---|
| 1. | **CLEAN**<br>Removes all nonprintable characters from text |
| 2. | **TRIM**<br>Removes spaces from text |

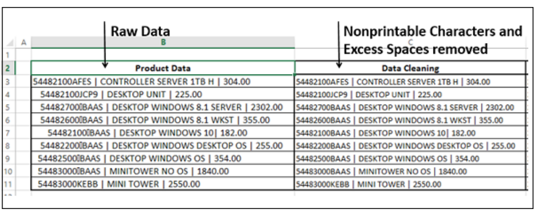
- Select the Cells C3 – C11.
- Type =TRIM (CLEAN (B3)) and then press CTRL + Enter.

The formula is filled in the cells C3 – C11.



| | A | B | C |
|---|---|---|---|
| 1 | | | |
| 2 | | **Product Data** | **Data Cleaning** |
| 3 | | 54482100AFES \| CONTROLLER SERVER 1TB H \| 304.00 | =TRIM(CLEAN(B3)) |
| 4 | | 54482100JCP9 \| DESKTOP UNIT \| 225.00 | =TRIM(CLEAN(B4)) |
| 5 | | 54482700□BAAS \| DESKTOP WINDOWS 8.1 SERVER \| 2302.00 | =TRIM(CLEAN(B5)) |
| 6 | | 54482600□BAAS \| DESKTOP WINDOWS 8.1 WKST \| 355.00 | =TRIM(CLEAN(B6)) |
| 7 | | 54482100□BAAS \| DESKTOP WINDOWS 10\| 182.00 | =TRIM(CLEAN(B7)) |
| 8 | | 54482200□BAAS \| DESKTOP WINDOWS DESKTOP OS \| 255.00 | =TRIM(CLEAN(B8)) |
| 9 | | 54482500□BAAS \| DESKTOP WINDOWS OS \| 354.00 | =TRIM(CLEAN(B9)) |
| 10 | | 54483000□BAAS \| MINITOWER NO OS \| 1840.00 | =TRIM(CLEAN(B10)) |
| 11 | | 54483000KEBB \| MINI TOWER \| 2550.00 | =TRIM(CLEAN(B11)) |

The result will be as shown below −



| | A | B | C |
|---|---|---|---|
| 1 | | | |
| 2 | | **Product Data** | **Data Cleaning** |
| 3 | | 54482100AFES \| CONTROLLER SERVER 1TB H \| 304.00 | 54482100AFES \| CONTROLLER SERVER 1TB H \| 304.00 |
| 4 | | 54482100JCP9 \| DESKTOP UNIT \| 225.00 | 54482100JCP9 \| DESKTOP UNIT \| 225.00 |
| 5 | | 54482700□BAAS \| DESKTOP WINDOWS 8.1 SERVER \| 2302.00 | 54482700BAAS \| DESKTOP WINDOWS 8.1 SERVER \| 2302.00 |
| 6 | | 54482600□BAAS \| DESKTOP WINDOWS 8.1 WKST \| 355.00 | 54482600BAAS \| DESKTOP WINDOWS 8.1 WKST \| 355.00 |
| 7 | | 54482100□BAAS \| DESKTOP WINDOWS 10\| 182.00 | 54482100BAAS \| DESKTOP WINDOWS 10\| 182.00 |
| 8 | | 54482200□BAAS \| DESKTOP WINDOWS DESKTOP OS \| 255.00 | 54482200BAAS \| DESKTOP WINDOWS DESKTOP OS \| 255.00 |
| 9 | | 54482500□BAAS \| DESKTOP WINDOWS OS \| 354.00 | 54482500BAAS \| DESKTOP WINDOWS OS \| 354.00 |
| 10 | | 54483000□BAAS \| MINITOWER NO OS \| 1840.00 | 54483000BAAS \| MINITOWER NO OS \| 1840.00 |
| 11 | | 54483000KEBB \| MINI TOWER \| 2550.00 | 54483000KEBB \| MINI TOWER \| 2550.00 |

### Finding required Text Patterns with the Text Functions

To structure your data, you might have to do certain Text Pattern matching based on which you can extract the Data Values. Some of the Text Functions that are useful for this purpose are −

| S.No. | Function & Description |
|---|---|
| 1. | **EXACT**<br>Checks to see if two text values are identical |
| 2. | **FIND**<br>Finds one text value within another (case-sensitive) |
| 3. | **SEARCH**<br>Finds one text value within another (not case-sensitive) |

## Extracting Data Values from Text

You need to extract the required data from text in order to structure the same. In the above example, say, you need to place the data in three columns – ProductID, Product_Description and Price.

You can extract data in one of the following ways −

- Extracting Data Values with Convert Text to Columns Wizard
- Extracting Data Values with Text Functions
- Extracting Data Values with Flash Fill

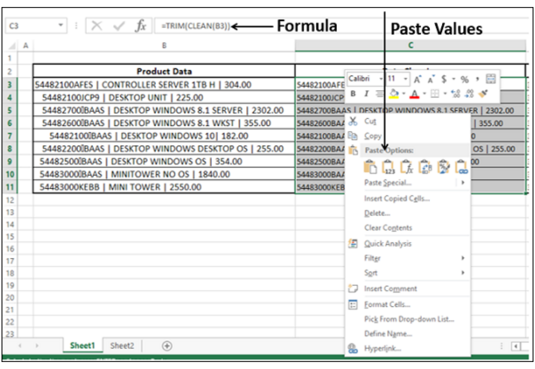
### Extracting Data Values with Convert Text to Columns Wizard

You can use the **Convert Text to Columns Wizard** to extract Data Values into Excel columns if your fields are −

- Delimited by a character, or
- Aligned in columns with spaces between each field.

In the above example, the fields are delimited by the character "|". Hence, you can use the **Convert Text to Columns** wizard.

- Select the data.
- Copy and paste values in the same place. Otherwise, **Convert Text to Columns** takes the functions rather than the data itself as the input.



=TRIM(CLEAN(B3))

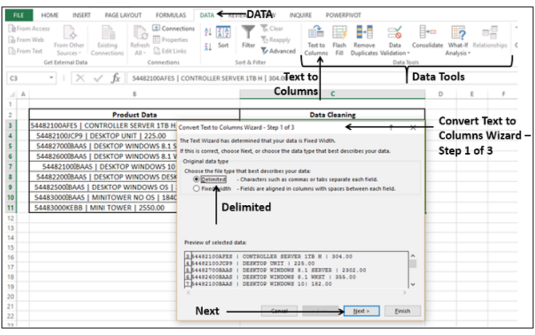
| | B |
|---|---|
| 2 | **Product Data** |
| 3 | 54482100AFES \| CONTROLLER SERVER 1TB H \| 304.00 |
| 4 | 54482100JCP9 \| DESKTOP UNIT \| 225.00 |
| 5 | 54482700□BAAS \| DESKTOP WINDOWS 8.1 SERVER \| 2302.00 |
| 6 | 54482600□BAAS \| DESKTOP WINDOWS 8.1 WKST \| 355.00 |
| 7 | 54482100□BAAS \| DESKTOP WINDOWS 10\| 182.00 |
| 8 | 54482200□BAAS \| DESKTOP WINDOWS DESKTOP OS \| 255.00 |
| 9 | 54482500□BAAS \| DESKTOP WINDOWS OS \| 354.00 |
| 10 | 54483000□BAAS \| MINITOWER NO OS \| 1840.00 |
| 11 | 54483000KEBB \| MINI TOWER \| 2550.00 |

- Select the data.
- Click on **Text to Columns** in the **Data Tools** group under **Data** Tab on the Ribbon.

**Step 1** − Convert Text to Columns Wizard - Step 1 of 3 appears.

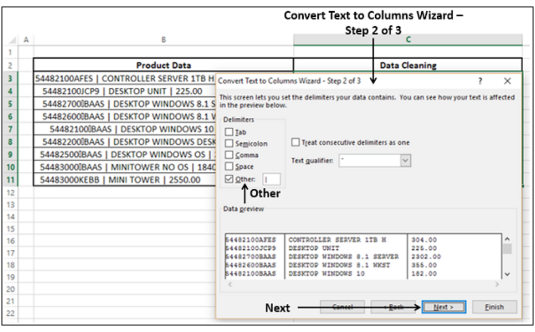
- Select Delimited.

- Click Next.



**Step 2** − Convert Text to Columns Wizard - Step 2 of 3 appears.

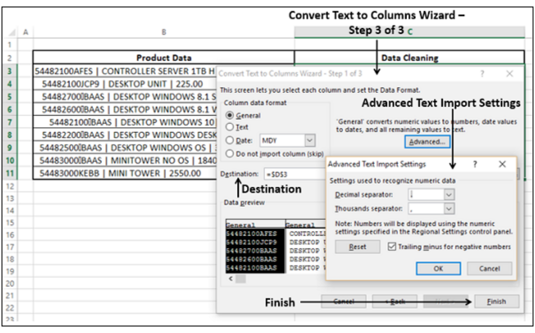
- Under **Delimiters**, select **Other**.
- In the box next to **Other**, type the character **|**
- Click **Next**.



**Step 3** − Convert Text to Columns Wizard - Step 3 of 3 appears.

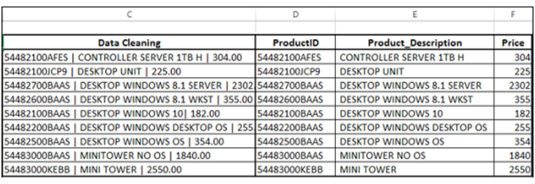
In this screen, you can select each column of your data in the wizard and set the format for that column.

- For **Destination**, select the cell D3.
- You can click **Advanced**, and set **Decimal Separator** and **Thousands Separator** in the **Advanced Text Import Settings** dialog box that appears.
- Click **Finish**.


Convert Text to Columns Wizard – Step 3 of 3
Advanced Text Import Settings
Destination
Finish

Your data, which is converted to columns appears in the three Columns – D, E and F.

- Name the Column headers as ProductID, Product_Description and Price.



| Data Cleaning | ProductID | Product_Description | Price |
|---|---|---|---|
| 54482100AFES \| CONTROLLER SERVER 1TB H \| 304.00 | 54482100AFES | CONTROLLER SERVER 1TB H | 304 |
| 54482100JCP9 \| DESKTOP UNIT \| 225.00 | 54482100JCP9 | DESKTOP UNIT | 225 |
| 54482700BAAS \| DESKTOP WINDOWS 8.1 SERVER \| 2302. | 54482700BAAS | DESKTOP WINDOWS 8.1 SERVER | 2302 |
| 54482600BAAS \| DESKTOP WINDOWS 8.1 WKST \| 355.00 | 54482600BAAS | DESKTOP WINDOWS 8.1 WKST | 355 |
| 54482100BAAS \| DESKTOP WINDOWS 10\| 182.00 | 54482100BAAS | DESKTOP WINDOWS 10 | 182 |
| 54482200BAAS \| DESKTOP WINDOWS DESKTOP OS \| 255. | 54482200BAAS | DESKTOP WINDOWS DESKTOP OS | 255 |
| 54482500BAAS \| DESKTOP WINDOWS OS \| 354.00 | 54482500BAAS | DESKTOP WINDOWS OS | 354 |
| 54483000BAAS \| MINITOWER NO OS \| 1840.00 | 54483000BAAS | MINITOWER NO OS | 1840 |
| 54483000KEBB \| MINI TOWER \| 2550.00 | 54483000KEBB | MINI TOWER | 2550 |

## Extracting Data Values with Text Functions

Suppose the fields in your data neither are delimited by a character nor are aligned in columns with spaces between each field, you can use text functions to extract data values. Even in the case the fields are delimited, you can still use text functions to extract data.

Some of the text functions that are useful for this purpose are −

| S.No. | Function & Description |
|---|---|
| 1. | **LEFT**<br>Returns the leftmost characters from a text value |
| 2. | **RIGHT**<br>Returns the rightmost characters from a text value |
| 3. | **MID**<br>Returns a specific number of characters from a text string starting at the position you specify |
| 4. | **LEN**<br>Returns the number of characters in a text string |

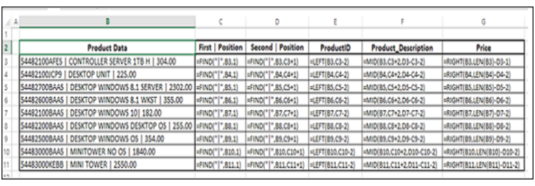
You can also combine two or more of these text functions as per the data you have at hand, to extract the required data values. For example, using a combination of LEFT, RIGHT and VALUE functions or using a combination of FIND, LEFT, LEN and MID functions.

In the above example,

- All the characters left to the first | give the name ProductID.
- All the characters right to the second | give the name Price.
- All the characters that lie between the first | and second | give the name Product_Description.
- Each | has a space before and after.

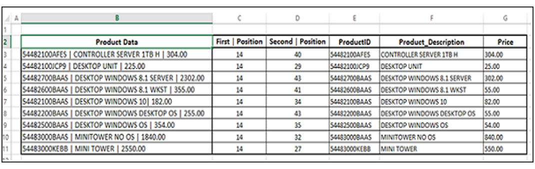
Observing this information, you can extract the data values with the following steps −

- Find the Position of First | - **First | Position**
  - You can use FIND function
- Find the Position of Second | - **Second | Position**
  - You can use FIND function again
- Beginning to (**First | Position** – 2) Characters of the Text give ProductID
  - You can use LEFT Function
- (**First | Position** + 2) to (**Second | Position** - 2) Characters of the Text give Product_Description
  - You can use MID Function
- (**Second | Position** + 2) to End Characters of the Text give Price
  - You can use RIGHT Function



| Product Data | First \| Position | Second \| Position | ProductID | Product_Description | Price |
|---|---|---|---|---|---|
| 54482100AFES \| CONTROLLER SERVER 1TB H \| 304.00 | =FIND("\|",B3,1) | =FIND("\|",B3,C3+1) | =LEFT(B3,C3-2) | =MID(B3,C3+2,D3-C3-2) | =RIGHT(B3,LEN(B3)-D3-1) |
| 54482100JCP9 \| DESKTOP UNIT \| 225.00 | =FIND("\|",B4,1) | =FIND("\|",B4,C4+1) | =LEFT(B4,C4-2) | =MID(B4,C4+2,D4-C4-2) | =RIGHT(B4,LEN(B4)-D4-2) |
| 54482700BAAS \| DESKTOP WINDOWS 8.1 SERVER \| 2302.00 | =FIND("\|",B5,1) | =FIND("\|",B5,C5+1) | =LEFT(B5,C5-2) | =MID(B5,C5+2,D5-C5-2) | =RIGHT(B5,LEN(B5)-D5-2) |
| 54482600BAAS \| DESKTOP WINDOWS 8.1 WKST \| 355.00 | =FIND("\|",B6,1) | =FIND("\|",B6,C6+1) | =LEFT(B6,C6-2) | =MID(B6,C6+2,D6-C6-2) | =RIGHT(B6,LEN(B6)-D6-2) |
| 54482100BAAS \| DESKTOP WINDOWS 10\| 182.00 | =FIND("\|",B7,1) | =FIND("\|",B7,C7+1) | =LEFT(B7,C7-2) | =MID(B7,C7+2,D7-C7-2) | =RIGHT(B7,LEN(B7)-D7-2) |
| 54482200BAAS \| DESKTOP WINDOWS DESKTOP OS \| 255.00 | =FIND("\|",B8,1) | =FIND("\|",B8,C8+1) | =LEFT(B8,C8-2) | =MID(B8,C8+2,D8-C8-2) | =RIGHT(B8,LEN(B8)-D8-2) |
| 54482500BAAS \| DESKTOP WINDOWS OS \| 354.00 | =FIND("\|",B9,1) | =FIND("\|",B9,C9+1) | =LEFT(B9,C9-2) | =MID(B9,C9+2,D9-C9-2) | =RIGHT(B9,LEN(B9)-D9-2) |
| 54483000BAAS \| MINITOWER NO OS \| 1840.00 | =FIND("\|",B10,1) | =FIND("\|",B10,C10+1) | =LEFT(B10,C10-2) | =MID(B10,C10+2,D10-C10-2) | =RIGHT(B10,LEN(B10)-D10-2) |
| 54483000KEBB \| MINI TOWER \| 2550.00 | =FIND("\|",B11,1) | =FIND("\|",B11,C11+1) | =LEFT(B11,C11-2) | =MID(B11,C11+2,D11-C11-2) | =RIGHT(B11,LEN(B11)-D11-2) |

The result will be as shown below −



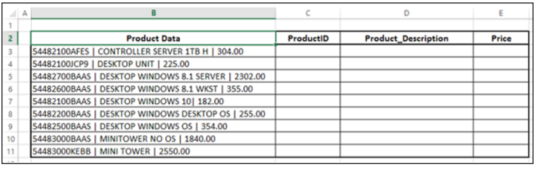
| Product Data | First \| Position | Second \| Position | ProductID | Product_Description | Price |
|---|---|---|---|---|---|
| 54482100AFES \| CONTROLLER SERVER 1TB H \| 304.00 | 14 | 40 | 54482100AFES | CONTROLLER SERVER 1TB H | 304.00 |
| 54482100JCP9 \| DESKTOP UNIT \| 225.00 | 14 | 29 | 54482100JCP9 | DESKTOP UNIT | 25.00 |
| 54482700BAAS \| DESKTOP WINDOWS 8.1 SERVER \| 2302.00 | 14 | 43 | 54482700BAAS | DESKTOP WINDOWS 8.1 SERVER | 302.00 |
| 54482600BAAS \| DESKTOP WINDOWS 8.1 WKST \| 355.00 | 14 | 41 | 54482600BAAS | DESKTOP WINDOWS 8.1 WKST | 55.00 |
| 54482100BAAS \| DESKTOP WINDOWS 10\| 182.00 | 14 | 34 | 54482100BAAS | DESKTOP WINDOWS 10 | 82.00 |
| 54482200BAAS \| DESKTOP WINDOWS DESKTOP OS \| 255.00 | 14 | 43 | 54482200BAAS | DESKTOP WINDOWS DESKTOP OS | 55.00 |
| 54482500BAAS \| DESKTOP WINDOWS OS \| 354.00 | 14 | 35 | 54482500BAAS | DESKTOP WINDOWS OS | 54.00 |
| 54483000BAAS \| MINITOWER NO OS \| 1840.00 | 14 | 32 | 54483000BAAS | MINITOWER NO OS | 840.00 |
| 54483000KEBB \| MINI TOWER \| 2550.00 | 14 | 27 | 54483000KEBB | MINI TOWER | 550.00 |

You can observe that the values in the price column are text values. To perform calculations on these values, you have to format the corresponding cells. You can look at the section given below to understand formatting text.

## Extracting Data Values with Flash Fill

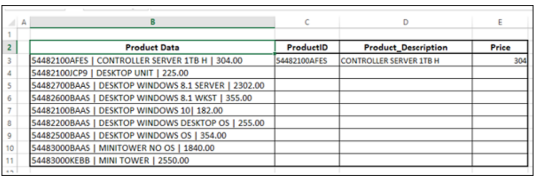
Using Excel **Flash Fill** is another way to extract data values from text. However, this works only when Excel is able to find a pattern in the data.

**Step 1** − Create three columns for ProductID, Product_Description and Price next to the data.



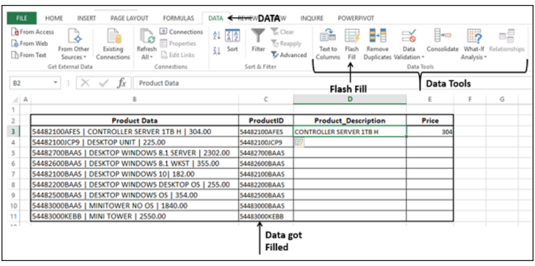
| Product Data | ProductID | Product_Description | Price |
|---|---|---|---|
| 54482100AFES \| CONTROLLER SERVER 1TB H \| 304.00 | | | |
| 54482100JCP9 \| DESKTOP UNIT \| 225.00 | | | |
| 54482700BAAS \| DESKTOP WINDOWS 8.1 SERVER \| 2302.00 | | | |
| 54482600BAAS \| DESKTOP WINDOWS 8.1 WKST \| 355.00 | | | |
| 54482100BAAS \| DESKTOP WINDOWS 10\| 182.00 | | | |
| 54482200BAAS \| DESKTOP WINDOWS DESKTOP OS \| 255.00 | | | |
| 54482500BAAS \| DESKTOP WINDOWS OS \| 354.00 | | | |
| 54483000BAAS \| MINITOWER NO OS \| 1840.00 | | | |
| 54483000KEBB \| MINI TOWER \| 2550.00 | | | |

**Step 2** − Copy and paste the values for C3, D3 and E3 from B3.



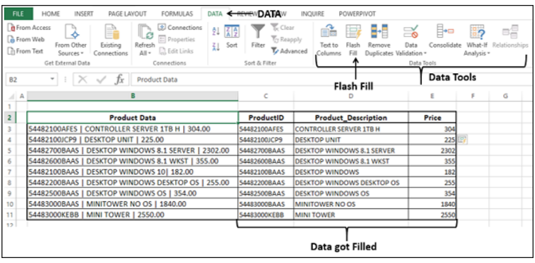
| Product Data | ProductID | Product_Description | Price |
|---|---|---|---|
| 54482100AFES \| CONTROLLER SERVER 1TB H \| 304.00 | 54482100AFES | CONTROLLER SERVER 1TB H | 304 |
| 54482100JCP9 \| DESKTOP UNIT \| 225.00 | | | |
| 54482700BAAS \| DESKTOP WINDOWS 8.1 SERVER \| 2302.00 | | | |
| 54482600BAAS \| DESKTOP WINDOWS 8.1 WKST \| 355.00 | | | |
| 54482100BAAS \| DESKTOP WINDOWS 10\| 182.00 | | | |
| 54482200BAAS \| DESKTOP WINDOWS DESKTOP OS \| 255.00 | | | |
| 54482500BAAS \| DESKTOP WINDOWS OS \| 354.00 | | | |
| 54483000BAAS \| MINITOWER NO OS \| 1840.00 | | | |
| 54483000KEBB \| MINI TOWER \| 2550.00 | | | |

**Step 3** − Select cell C3 and click **Flash Fill** in the **Data Tools** group on the **Data** tab. All the values for ProductID get filled.



| Product Data | ProductID | Product_Description | Price |
|---|---|---|---|
| 54482100AFES \| CONTROLLER SERVER 1TB H \| 304.00 | 54482100AFES | CONTROLLER SERVER 1TB H | 304 |
| 54482100JCP9 \| DESKTOP UNIT \| 225.00 | 54482100JCP9 | | |
| 54482700BAAS \| DESKTOP WINDOWS 8.1 SERVER \| 2302.00 | 54482700BAAS | | |
| 54482600BAAS \| DESKTOP WINDOWS 8.1 WKST \| 355.00 | 54482600BAAS | | |
| 54482100BAAS \| DESKTOP WINDOWS 10\| 182.00 | 54482100BAAS | | |
| 54482200BAAS \| DESKTOP WINDOWS DESKTOP OS \| 255.00 | 54482200BAAS | | |
| 54482500BAAS \| DESKTOP WINDOWS OS \| 354.00 | 54482500BAAS | | |
| 54483000BAAS \| MINITOWER NO OS \| 1840.00 | 54483000BAAS | | |
| 54483000KEBB \| MINI TOWER \| 2550.00 | 54483000KEBB | | |

**Step 4** − Repeat the above given steps for Product_Description and Price. The data is filled.



| Product Data | ProductID | Product_Description | Price |
|---|---|---|---|
| 54482100AFES \| CONTROLLER SERVER 1TB H \| 304.00 | 54482100AFES | CONTROLLER SERVER 1TB H | 304 |
| 54482100JCP9 \| DESKTOP UNIT \| 225.00 | 54482100JCP9 | DESKTOP UNIT | 225 |
| 54482700BAAS \| DESKTOP WINDOWS 8.1 SERVER \| 2302.00 | 54482700BAAS | DESKTOP WINDOWS 8.1 SERVER | 2302 |
| 54482600BAAS \| DESKTOP WINDOWS 8.1 WKST \| 355.00 | 54482600BAAS | DESKTOP WINDOWS 8.1 WKST | 355 |
| 54482100BAAS \| DESKTOP WINDOWS 10\| 182.00 | 54482100BAAS | DESKTOP WINDOWS | 182 |
| 54482200BAAS \| DESKTOP WINDOWS DESKTOP OS \| 255.00 | 54482200BAAS | DESKTOP WINDOWS DESKTOP OS | 255 |
| 54482500BAAS \| DESKTOP WINDOWS OS \| 354.00 | 54482500BAAS | DESKTOP WINDOWS OS | 354 |
| 54483000BAAS \| MINITOWER NO OS \| 1840.00 | 54483000BAAS | MINITOWER NO OS | 1840 |
| 54483000KEBB \| MINI TOWER \| 2550.00 | 54483000KEBB | MINI TOWER | 2550 |

## Formatting Data with Text Functions

Excel has several built-in text functions that you can use for formatting data containing text. These include −

**Functions that format the Text as per your need** −

| S.No. | Function & Description |
|---|---|
| 1. | **LOWER**<br>Converts text to lowercase |

| S.No. | Function & Description |
|---|---|
| 1. | **UPPER**<br>Converts text to uppercase |
| 2. | **PROPER**<br>Capitalizes the first letter in each word of a text value |

**Functions that convert and/or format the Numbers as Text** −

| S.No. | Function & Description |
|---|---|
| 1. | **DOLLAR**<br>Converts a number to text, using the $ (dollar) currency format |
| 2. | **FIXED**<br>Formats a number as text with a fixed number of decimals |
| 3. | **TEXT**<br>Formats a number and converts it to text |

**Functions that convert the Text to Numbers** −

| S.No. | Function & Description |
|---|---|
| 1. | **VALUE**<br>Converts a text argument to a number |

**Executing Data Operations with the Text Functions**

You might have to perform certain Text Operations on your Data. For example, if Login-IDs for the Employees are changed to a New Format in an Organization, based on the Format Change, Text Replacements might have to be done.

Following Text Functions help you in performing Text Operations on your data containing Text −

| S.No. | Function & Description |
|---|---|
| 1. | **REPLACE**<br>Replaces characters within text |
| 2. | **SUBSTITUTE**<br>Substitutes new text for old text in a text string |
| 3. | **CONCATENATE**<br>Joins several text items into one text item |
| 4. | **CONCAT**<br>Combines the text from multiple ranges and/or strings, but it does not provide the delimiter or IgnoreEmpty arguments. |
| 5. | **TEXTJOIN**<br>Combines the text from multiple ranges and/or strings, and includes a delimiter you specify between each text value that will be combined. If the delimiter is an empty text string, this function will effectively concatenate the ranges. |
| 6. | **REPT**<br>Repeats text a given number of times |