Contents

[Cleaning & Transforming Data in Excel 1](#_Toc172795692)

[Overview 1](#_Toc172795693)

[Dataset 1](#_Toc172795694)

[Step 1: Open Data in Power Query in Excel 1](#_Toc172795695)

[Step 2: Capitalise Column Names 2](#_Toc172795696)

[Step 3: Identify and Replace Null Values in Ship Date 2](#_Toc172795697)

[Step 4: Fix Any Swapped Order Dates and Ship Dates 4](#_Toc172795698)

[Step 5: Identify and Replace Null Values in Country 5](#_Toc172795699)

[Step 6: Split City Column into City and State columns 6](#_Toc172795700)

[Step 6: Remove alternative spellings from State column 6](#_Toc172795701)

[Step 7: Check and Correct Datatypes for Columns 6](#_Toc172795702)

[Step 8: Create New Calculated Columns 7](#_Toc172795703)

[Step 9: Load the Cleaned and Transformed Data into Excel 7](#_Toc172795704)

[Questions? 7](#_Toc172795705)

# Cleaning & Transforming Data in Excel

## Overview

In this activity, you will practice cleaning and transforming a dataset in Power Query in Excel. You will work with a modified Superstore dataset that contains various data issues. Your goal is to clean the data so that we can carry out an accurate analysis.

## Dataset

sample\_-\_superstore.xlsx

## Step 1: Open Data in Power Query in Excel

1. Open a blank worksheet in Excel.
2. Under the *Data* tab at the top of the screen, click on **Get Data**.
3. Select **Excel** worksheet and navigate to sample superstore datafile.
4. In the navigator, choose the **Orders** table on the left
5. Select **Transform Data** on the bottom of the navigator window.

## Step 2: Capitalise Column Names

1. **Scan** the **Header Row** to see what columns are in the dataset and what needs to be formatted correctly.
2. **Right click** on *region* column header andSelect **Rename**
3. **Rename column** to capitalised name,*Region***.**
4. **Repeat these steps** for the other columns that are not capitalised: *category* and*sales***.**

## Step 3: Identify and Replace Null Values in Ship Date

1. **View Column Profile**
   * Select View tab at top of screen.
   * Check the Column Profile box in the Data Preview group.
   * Click on the Row ID column header.
   * Looking at the column statistics, note that there are no empty or missing values in the Row ID column.
   * Click on the column headers for the Order ID and Order Date columns as well and note that they also have no empty or missing values.
   * Click on the Ship Date column header.
   * Note that Ship Date has 1 empty value.

A screenshot of a computer

Description automatically generated

* + Uncheck column profile in the data preview group to free room on the screen.

1. Filter for null values in *Ship Date* column
   * Click the down arrow in the Ship Date column header
   * Uncheck Select All and Check Null.
   * Click OK.
2. Let’s see what the ship date is for other items in the same order.
   * Select the Order ID in the row with the null ship date and copy that order ID from the bottom of the screen. (click and drag > CTRL-C)
   * Now remove the filter from the Ship Date column.
     1. Click the filter icon (with the funnel) in the column header
     2. Uncheck null and then check Select All.
     3. Click OK.
   * Filter the Order ID column for the Order ID you copied previously from the row with the null Ship Date.
     1. Click the down arrow in the Order ID column header
     2. Paste the copied Order ID (CA-2014-115812) into the Search field.
     3. Click OK.

A screenshot of a computer

Description automatically generated

1. **Infer Missing Ship Date**
   * Note that the ship date for that Order ID is 14/06/2014 for all the rows except the row with a null value and one row that seems to have the order date and ship date swapped.
   * First **undo the filtering** in the *Order ID* column by looking under Query Settings on the right of the screen and, under Applied Steps, checking the X to the left of Filtered Rows. *You can always undo any changes you make here in Applied Steps.*

*A screenshot of a computer

Description automatically generated*

* + Then **Replace the *null* value with the correct date, *14/06/2014*.**
    1. Right Click the *Ship Date* column header and select **Replace Values**.
    2. Enter *null* in the **Value to Find** field and enter *14/06/2014* in the **Replace With** field.
    3. Click **OK**.

## Step 4: Fix Any Swapped Order Dates and Ship Dates

It’s not possible for the ship date to be before the order date so this is a data validation issue. It doesn’t follow business logic. Let’s make sure we fix all instances where this happened.

1. **Create a New Order Date column**
   * Select **Add Column** tab.
   * Click on **Custom Column** in the general group.
   * Under New Column Name, enter New Order Date.
   * Under Custom Column Formula, after the equal sign, enter *if [Order Date] > [Ship Date] then [Ship Date] else [Order Date]*
   * Click OK.
2. **Create a New Ship Date column**
   * Select **Add Column** tab.
   * Click on **Custom Column** in the general group.
   * Under New Column Name, enter New Ship Date.
   * Under Custom Column Formula, after the equal sign, enter *if [Order Date] > [Ship Date] then [Order Date] else [Ship Date])*
   * Click OK.
3. **Delete Old Order Date and Ship Date columns**
   * Go to Home tab.
   * Click Choose Columns > Choose Columns in the manage columns group.
   * Uncheck Order Date and Ship Date.Click OK

A screenshot of a computer

Description automatically generated

1. **Move New Columns Closer to the Beginning of Table**
   * **Move the New Order Date and New Ship Date columns**
     1. Hold down CTRL and Click the headers for both new columns, *New Order Date* and *New Ship Date*, to select them.
     2. Right Click either header and select *Move > To Beginning*.
   * **Move the Row ID and Order ID columns to Start**
     1. Hold down CTRL and Click the headers for *Row ID* and *Order ID* columns to select them.
     2. Right Click either header and select *Move > To Beginning*.
   * **Rename New Order Date and New Ship Date columns**
     1. Right Click on column header for New Order Date and Rename to *Order Date*.
     2. Do the same to rename New Ship Date to *Ship Date*.

## Step 5: Identify and Replace Null Values in Country

1. **View Column Profile**
   * Select View tab at top of screen.
   * Check the Column Profile box in the Data Preview group.
   * Click on each of the column headers until, looking at the column statistics, you find one that has empty or missing values. We need to remedy the 5 missing values in the country column.
   * Look at the value distribution in the Country column and note that all the orders are from the United States. Therefore we can replace any missing values in the Country column with United States.

A screenshot of a computer

Description automatically generated

* + Uncheck Column Profile to free up screen space.
  + Right Click on Country column and select Replace Values.
  + Replace null with United States.

## Step 6: Split City Column into City and State columns

1. **Right Click on the City Column**
2. Go to the Transform Tab at the top of the screen.
3. **Select Split Column > By Lowercase to Uppercase**
4. Rename the Split Columns
   * Rename the left split column, City.1, to City.
   * Rename the right split column, City.2, to State.

## Step 6: Remove alternative spellings from State column

1. Click the down arrow in the State header and scroll through the alphabetical list of states in the column to see that New York is in there twice, spelled as New York and NY. We need to replace the NY with New York.
2. Right click on the header for the State column and Select Replace Values.
3. **Value to Find**: NY, **Replace With**: New York.
4. Click **OK.**

## Step 7: Check and Correct Datatypes for Columns

1. Click the icon to the left of each column header to see what datatype is selected for that column.
2. If the datatype does not match the contents of the column then we have to select the correct datatype from the list.
   * Change the datatype for the *Profit* column to **currency**.

**A screenshot of a computer

Description automatically generated**

* + Change the Order Date and Ship Date columns to date datatype.

## Step 8: Create New Calculated Columns

1. **Calculate Processing Time (Optional Challenge Exercise):**
   * Select the Add Column Tab.
   * Click on Custom Column.
   * Give the Column a name of Processing Time.
   * Write a formula for the difference between Ship Date and Order Date.
   * Click OK.
   * Change datatype for column to Duration.

## Step 9: Load the Cleaned and Transformed Data into Excel

1. Select the Home tab.
2. Click Close and Apply to load the data into Excel.

Congratulations, you are done cleaning and transforming your data. This step can be as much as 40% of the work of a data analyst.

## Questions?

* If you have any questions or need assistance, feel free to ask your instructor or a peer.