Excel Data Model Cleaning Data- I

Dr. Aghabi Abosaif

Introduction

- It’s very common when collecting or importing data - whether through manual or automated processes - to **get errors** and **inconsistencies** in your data.

- This can be as simple as spelling mistakes, extra white space, or the wrong case used in text, empty rows or missing values in your data, to inaccurate or duplicated data.

- Having these **errors and inconsistencies** in your data can lead to **unsuccessful** sorting and filtering operations and therefore inadequately visualized and presented data findings.

Cleaning Data

- Understand how to deal with **irrelevant** or **inaccurate** data.
  - **Check spelling.**
  - Remove **empty rows**.
  - Remove **duplicated** data.
  - Change **text case.**
  - Change **date formatting.**

1. Check Spelling

- Open Dataset (CustomerOrders-01)
- Select specific column (j (ShipCity)), then click **Review** tab, and select **Spelling**.
  - Click the correct suggestion to change the spelling.
  - Close the **Spelling** pane.

- ***Remember check spelling must do only in columns that contain a dictionary words***

2. Remove Empty Rows

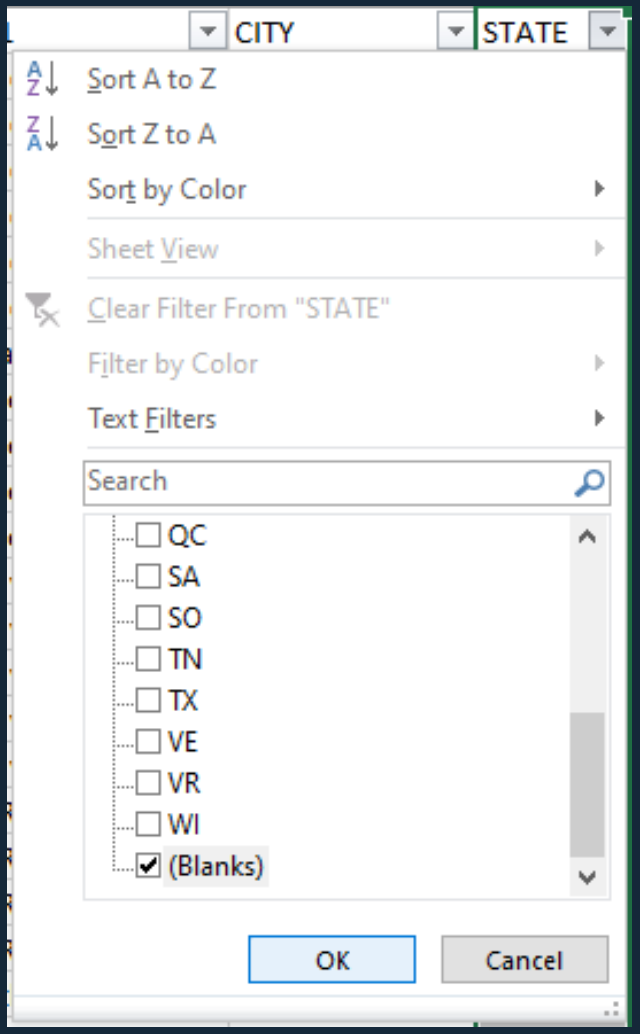
- CTRL+DOWN ARROW note what happen?
- The cursor keeps **stopping** when it gets to an **empty row** meaning that the dataset if essentially being split into multiple sections, separated by these empty rows.

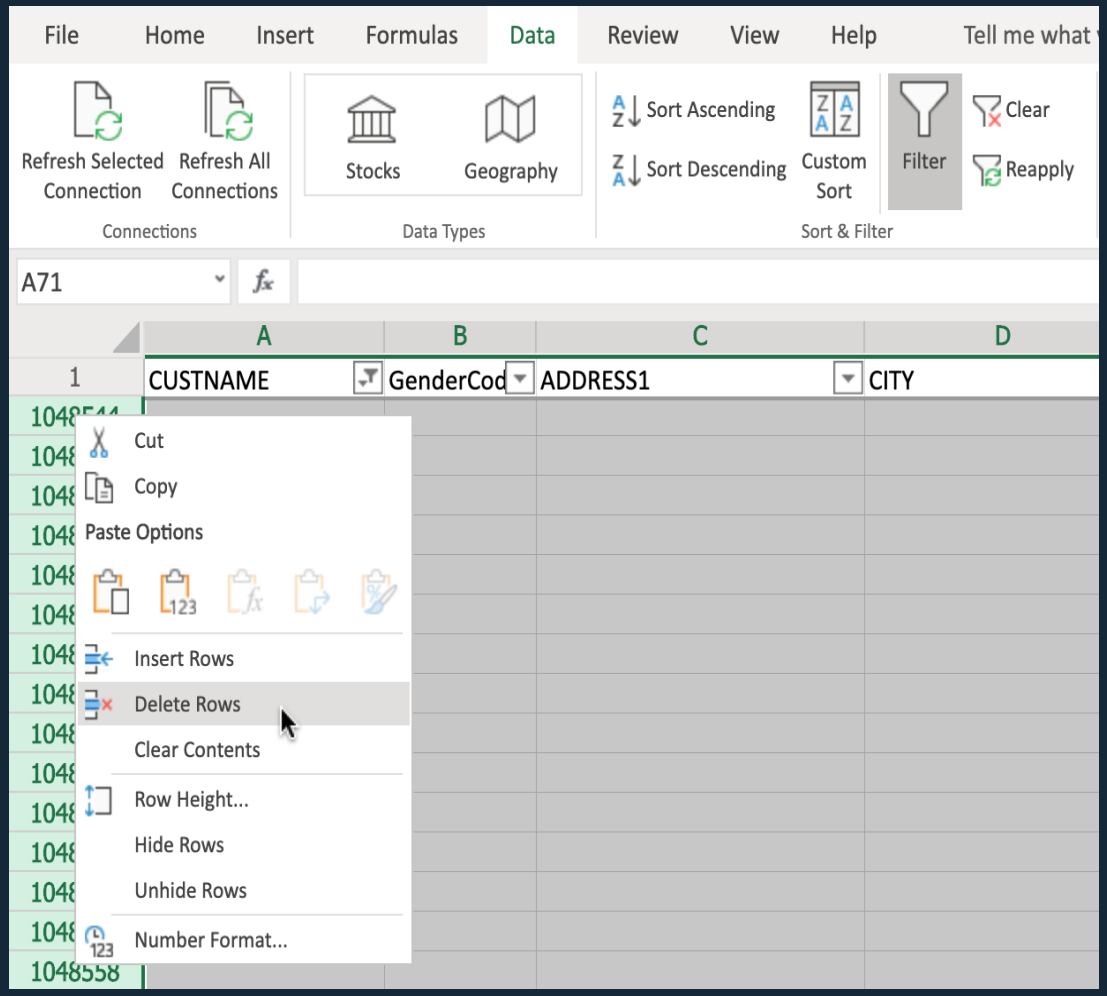
- Two options; *one option* is to just Manually scroll down the sheet looking for empty rows and deleting each one.

2. Remove Empty Rows

- Second Option:
- Press **CTRL+HOME**, then press **CTRL+SHIFT+END** to select the whole datasheet.
- On the **Data** tab, click **Filter**.
- Press **CTRL+HOME**, click the **filter arrow** in the **ShippedDate** column, and then click **Filter**.
- Click the **Select All** checkbox to deselect all of them. Then select just **Blanks**, then **OK**.
- Select **first row**, then press **CTRL+SHIFT+END** to select all rows.
- Right-click the selected rows and then click **Delete Rows**.
- Finally, on the **Data** tab, click **Clear**, then click **Filter**.

2. Remove Empty Rows

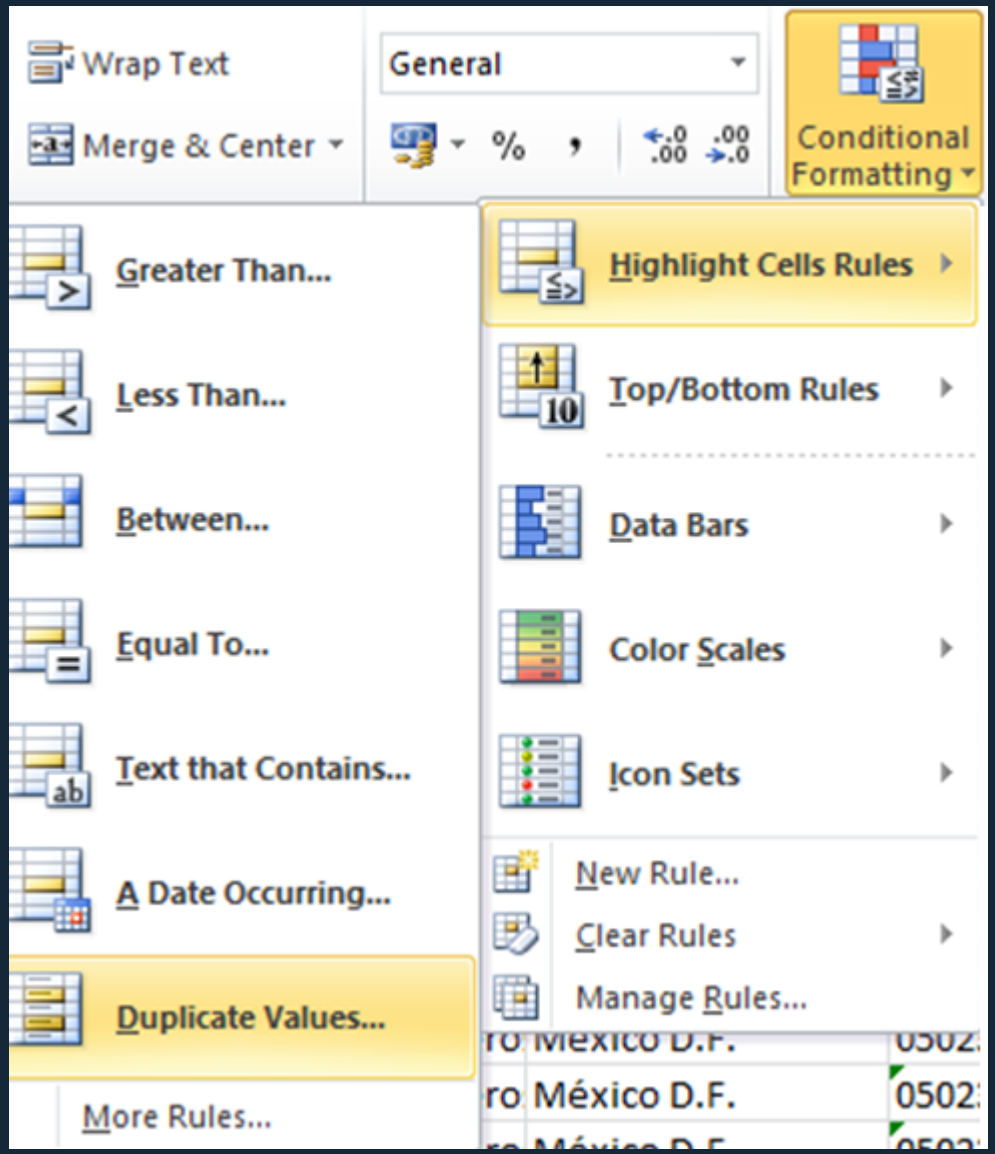




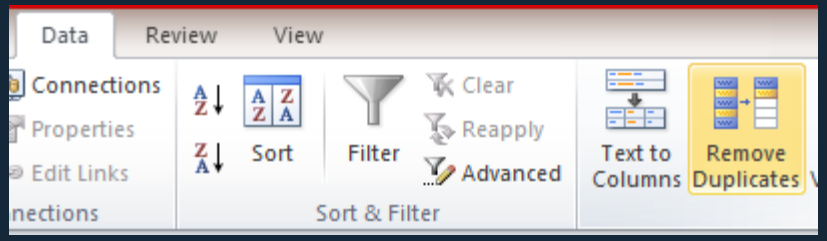
3. Remove Duplicate Rows

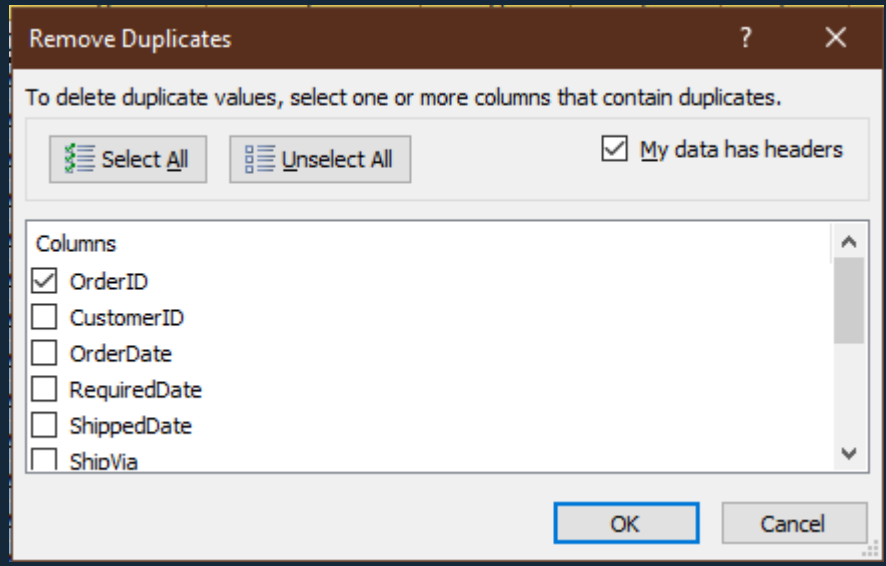
- You can **Highlight the Duplicated values** first then decided to remove it or no.
- Select Column **A (OrderID Or ShipName)** since ORDER_ID values **are unique**.
- On the **Home** tab, click **Conditional Formatting> Highlight Cells Rules> Duplicate Values**, and then click **OK**.
- Select the whole datasheet **(CTRL+SHIFT+END)**
- On the **Data** tab, click **Remove Duplicates**.
- 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**.
- In the pop-up box informing you how many duplicate values were found and removed, click **OK**.

Highlight the Duplicated values



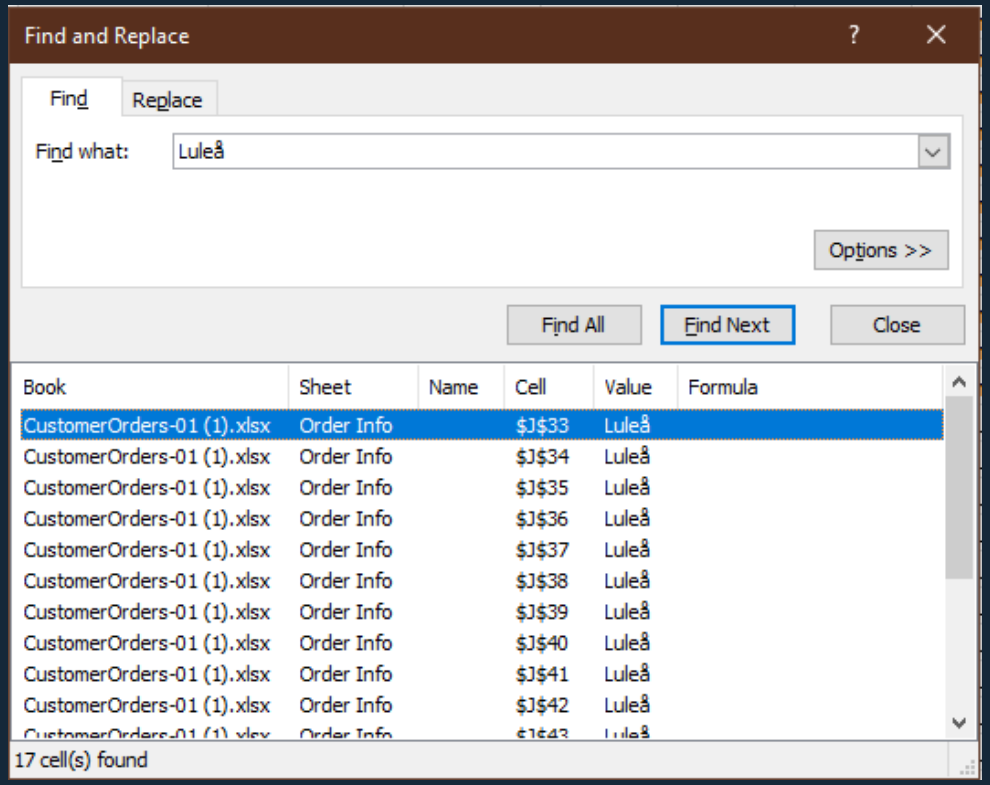
Remove the Duplicated values

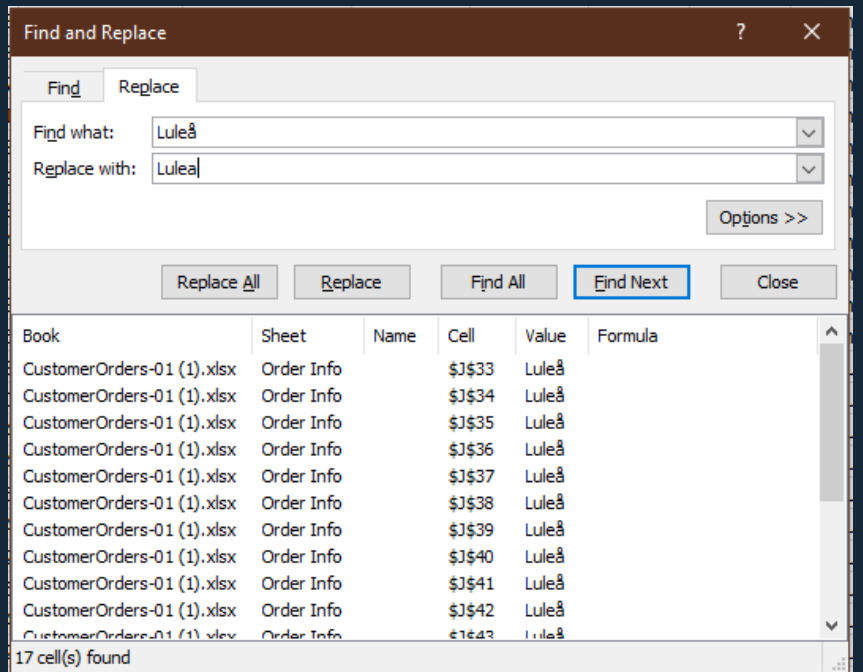




4. Use Find & Replace to Correct Misspelling

- On the **Home** tab, click **Find & Select**.
- Click **Find**. In Find what, type **Mexico**, and click **Find All**.
- Click **Replace**.
- In Replace with, type **Mexico.D.F**, click **Replace All**, and then click the **Close** icon.


Find and Replace
Find
Replace
Find what: Luleå
Options >>
Find All
Find Next
Close
Book Sheet Name Cell Value Formula
CustomerOrders-01 (1).xlsx Order Info J33 Luleå
CustomerOrders-01 (1).xlsx Order Info J34 Luleå
CustomerOrders-01 (1).xlsx Order Info J35 Luleå
CustomerOrders-01 (1).xlsx Order Info J36 Luleå
CustomerOrders-01 (1).xlsx Order Info J37 Luleå
CustomerOrders-01 (1).xlsx Order Info J38 Luleå
CustomerOrders-01 (1).xlsx Order Info J39 Luleå
CustomerOrders-01 (1).xlsx Order Info J40 Luleå
CustomerOrders-01 (1).xlsx Order Info J41 Luleå
CustomerOrders-01 (1).xlsx Order Info J42 Luleå
17 cell(s) found


Find and Replace
Find
Replace
Find what: Luleå
Replace with: Lulea
Options >>
Replace All
Replace
Find All
Find Next
Close
Book Sheet Name Cell Value Formula
CustomerOrders-01 (1).xlsx Order Info J33 Luleå
CustomerOrders-01 (1).xlsx Order Info J34 Luleå
CustomerOrders-01 (1).xlsx Order Info J35 Luleå
CustomerOrders-01 (1).xlsx Order Info J36 Luleå
CustomerOrders-01 (1).xlsx Order Info J37 Luleå
CustomerOrders-01 (1).xlsx Order Info J38 Luleå
CustomerOrders-01 (1).xlsx Order Info J39 Luleå
CustomerOrders-01 (1).xlsx Order Info J40 Luleå
CustomerOrders-01 (1).xlsx Order Info J41 Luleå
CustomerOrders-01 (1).xlsx Order Info J42 Luleå
17 cell(s) found

5. CHOOSE() Function

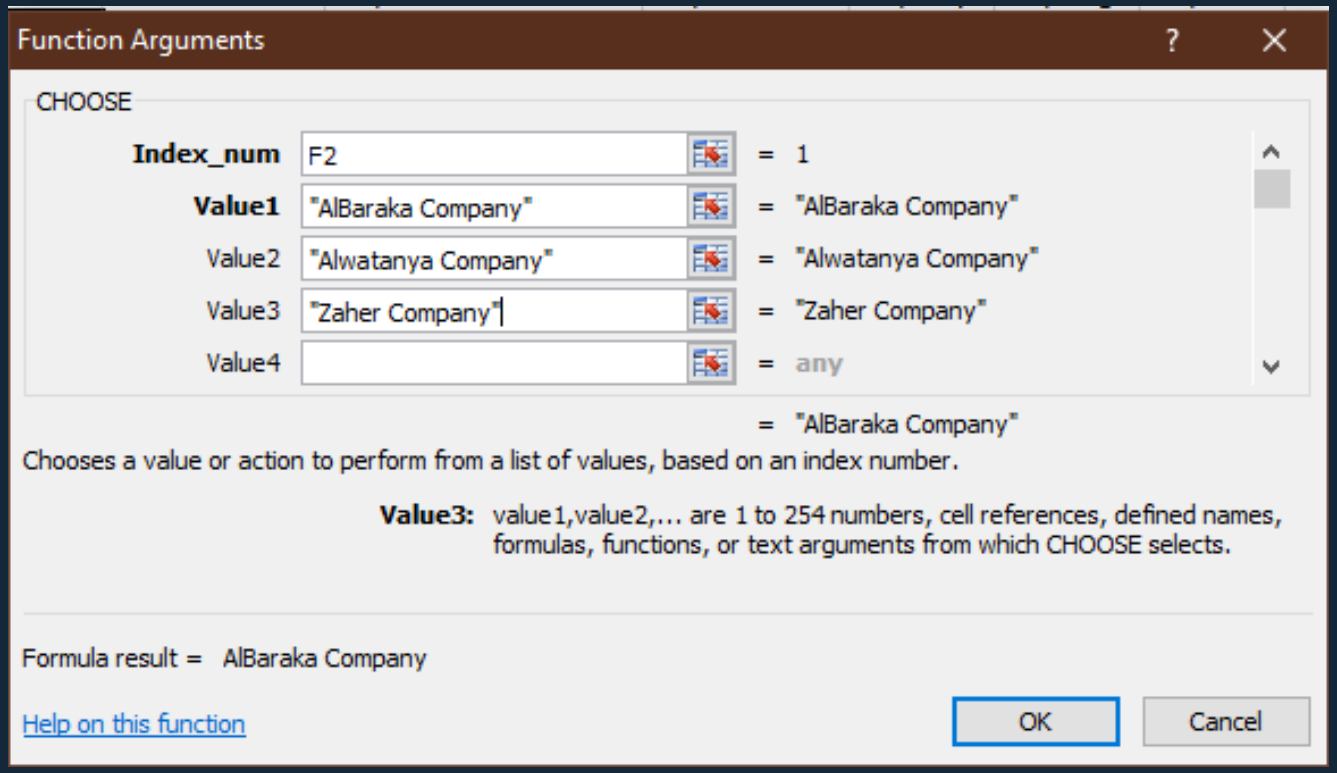
- It`s **similar** to find & replace Functions.
- It designed to **return a value from the list based on a specified position**.
- It replace **Meaningless** Data with Excels with something more presentable, something we could actually use within a report.
- It's known as **an array function.**

CHOOSE() Function

- In Column **G(OrderAmount)** Press (Ctrl + '+')
- In Header Write '**ShipperName'.**
- In **G2** , write Choose Function as Shown Below.

CHOOSE() Function

=CHOOSE(F2,"AlBaraka Company","Alwatanya Company","Zaher Company")



Thank You For Your Attention