Data Cleaning Strategies in Excel

Malcolm G. Keif, Ph.D.

1. **Nothing is better** than validating data integrity on entry. Structured data validation using “tidy” fields (values comprised of one observation and one variable) is clearly the best approach. However, if you must clean an untidy dataset in Excel, consider these strategies:

2. **Multiple fields in one column**
   a. **Single row + well-delimited or fixed-width**: Text-to-Columns

   b. **Multiple rows per record + well delimited/fixed width**: *=SUBSTITUTE(B2,CHAR(10),"")*
      i. CHAR(10) is the 10th character in ASCII, which is a line break.
      ii. Must then Copy and Paste *Values* in a new row to continue working with this data.

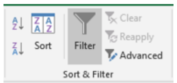
   c. **Not well delimited/variable width**: much more difficult – options include…
      i. Extracting Honorifics/Prefixes – Using **LEFT**:
      =IF(SUM(--ISNUMBER(SEARCH({"Ms.","Dr.","Mr.","Miss","Rev.","Prof.","Captain"}, F2)))>0,LEFT(F2,FIND(" ",F2)-1),"")
         1. {} represents an array that is used for searching in the field (F2). This approach requires every prefix option be included in your array.
         2. The double minus signs before ISNUMBER (--) change Boolean results to binary – 1,0 instead of True,False, making it so you can sum them.
         3. Sum is used to add the 1,0 for the array search. When greater than 0 (meaning a match), returns the Left word up to the space “ “. If 0, returns the closest thing Excel has to a null field (it is actually a text field with no value).
         4. LEFT returns the value from F2 starting from the left and returning a given number of characters, in this case after FINDing the first space, minus 1 to not include the space in the extraction.
         5. Alternately, you can name your array (using Name Box / Name Manger) and refer to the cell range in your formula, but you will need to use CTRL+SHIFT+ENTER to save it as an array formula (will display with brackets {}).
            a. =IF(SUM(--ISNUMBER(SEARCH(honorific,F2)))>0,LEFT(F2,FIND(" ",F2)-1),"")

      ii. Extracting Credentials/suffixes – Using SUBSTITUTE & TRIM when working from the RIGHT:
      =IF(SUM(--ISNUMBER(SEARCH({"MBA","Ph.D.","MD","DDS","Esq."},I2)))>0, TRIM(RIGHT(SUBSTITUTE(I2," ",REPT(" ", 99)), 99)),"")
         1. Similar to the above in the logic test to SEARCH for a match with the words in the array. This approach requires every suffix option be included in your array.
         2. When working from the right, FIND (used above) may not work because the function starts from the left. The work around it…
         3. This function SUBSTITUTEs all characters to the left of the RIGHT most word with 99 REPT (repeating) spaces, and then TRIMs the 99 spaces back down to a single space.

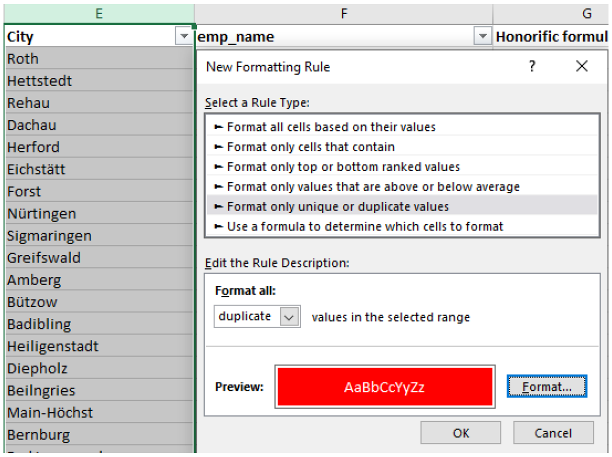
      iii. Using LEN (Length) as a way to indicate starting place for extraction:
      =IF(G2="",F2,RIGHT(F2,LEN(F2)-LEN(G2)-1))
         1. LEN is a useful way for calculating starting points for returning values with removed prefixes or suffixes.

      iv. Multiple given names (middle name or multiple first names) add significant complexity and may not be achievable without manual editing.

3. **Irrelevant records or outliers**
   a. Sorting and Filtering are often simple ways to identify outliers when the data set isn't too large or when fields have upper/lower limits or are nominal data.
   b. Select the top row, go to the DATA ribbon and select filter.
   c. You can sort from within the filter menu.
   d. If the dataset isn't too large, scroll down and look for outliers.
   e. Generally, we do not delete outliers but rather "exclude" them in analysis/visualization.



4. **Duplicate Records**
   a. Conditional Formatting is often a simple way to identify duplicates.
   b. The range of cells with which you are checking for duplicates must be selected before applying the conditional formatting.



E	F	G
City	emp_name	Honorific formul

Roth
Hettstedt
Rehau
Dachau
Herford
Eichstätt
Forst
Nürtingen
Sigmaringen
Greifswald
Amberg
Bützow
Badibling
Heiligenstadt
Diepholz
Beilngries
Main-Höchst
Bernburg

New Formatting Rule ? ×

Select a Rule Type:

- Format all cells based on their values
- Format only cells that contain
- Format only top or bottom ranked values
- Format only values that are above or below average
- Format only unique or duplicate values
- Use a formula to determine which cells to format

Edit the Rule Description:

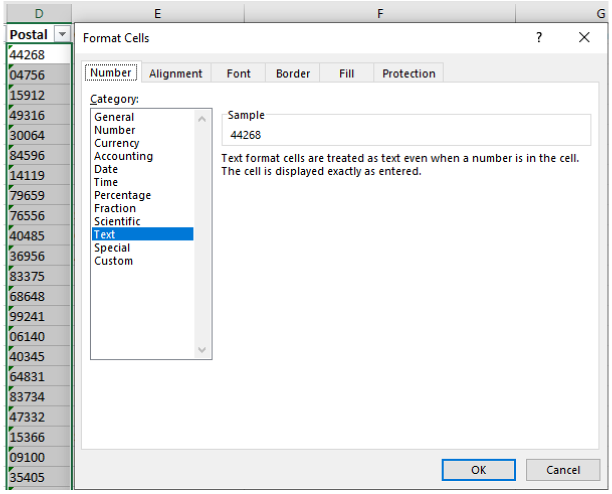
Format all:

duplicate values in the selected range

Preview: AaBbCcYyZz Format...

OK Cancel

5. **Converting data types**
   a. Select the column and then go to Cell Formatting.



D
Postal
44268
04756
15912
49316
30064
84596
14119
79659
76556
40485
36956
83375
68648
99241
06140
40345
64831
83734
47332
15366
09100
35405

Format Cells ? ×

Number | Alignment | Font | Border | Fill | Protection

Category:
General
Number
Currency
Accounting
Date
Time
Percentage
Fraction
Scientific
Text
Special
Custom

Sample
44268

Text format cells are treated as text even when a number is in the cell. The cell is displayed exactly as entered.

OK Cancel

6. Finally, you may use *Find and Replace* to make global changes in your document. For example, if you want to get rid of honorifics ("Dr." for example), you may search for it (Find under Edit menu) in your data and replace it with no value, in effect deleting it. You want to be cautious though as it is preferred not to permanently delete your original data, so make sure you have a backup of the original. Alternately, you can place any fields you think you don't need into columns you won't use, preserving the data if needed.