FOSE1025 — Scientific Computing

Week 7 Lecture 1: Cleaning Data

Diego Mollá

FOSE1025 2020H2

Abstract

In this lecture we will revise the typical steps in a data science project and will focus in one of the first steps: data cleaning. We will look at various tools that Excel provides to help cleaning raw data: fixing spelling errors, change the data types, split columns, merge columns, and filling missing information. Much of the contents and examples in this lecture is based on the LinkedIn Learning course "Excel 2016: Cleaning up Your Data", which has an excellent collection of short videos that we recommend you should watch.

Update September 7, 2020

Contents

Reading

- LinkedIn Learning — Excel 2016: Cleaning up Your Data
  https://www.linkedin.com/learning/excel-2016-cleaning-up-your-data
- MATLAB Characters and Strings
  https://au.mathworks.com/help/matlab/characters-and-strings.html
- MATLAB for Data Processing and Visualization: Preprocessing Data
  https://matlabacademy.mathworks.com/R2020a/portal.html?course=mlvi

1 Cleaning Text Data in Excel

Text as Unstructured Data

- Much of the information you find is input in text.
- People can understand text very easily . . .
- . . . but not machines!
- Text is often called a kind of *unstructured data.*
- Excel and MATLAB can help find structure from text.

What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!
What they hear
blah blah GINGER blah blah blah blah blah blah blah blah GINGER blah blah blah blah blah blah...

Some Useful Text Functions

CH-05.xlsx From *https://www.linkedin.com/learning/excel-2016-cleaning-up-your-data/use-text-functions*

Name	Description
LOWER	Converts all text to lowercase
PROPER	Capitalizes only letters than start the entry or follow a space or punctuation
UPPER	Converts all text to uppercase
REPLACE	Replaces characters within text, based on content, not on character position
SUBSTITUTE	Replaces characters within text, based on character position, not on content
REPT	Repeats text a given number of times
LEFT	Returns the leftmost characters from a text value
MID	Returns a specific number of characters from a text string starting at the position you specify
RIGHT	Returns the rightmost characters from a text value
FIND	Finds one text value within another (case-sensitive)
SEARCH	Finds one text value within another (not case-sensitive)
EXACT	Checks to see if two text values are identical
LEN	Returns the number of characters in a text string
TEXT	Formats a number and converts it to text
VALUE	Converts a text argument to a number
CLEAN	Removes all nonprintable characters from text
TRIM	Removes spaces from text
CONCATENATE	Joins several text items into a cell (on older Excel versions)
CONCAT	Joins several text items into a cell (on newer Excel versions)
DOLLAR	Converts a number to text, using the $ (dollar) currency format
FIXED	Formats a number as text with a fixed number of decimals
TEXTJOIN	Joins several text items into a cell using a delimiter

Concatenating Text

Several ways to concatenate text:

- Using the & operator

```
=A1 & ” ” & B1
```

- CONCAT (in Excel versions from 2016, Mobile, Web)

```
=CONCAT(”Stream population for ”, A2, ” ”,
              A3, ” is ”, A4, ”/mile.”)
=CONCAT(B2:C8)
```

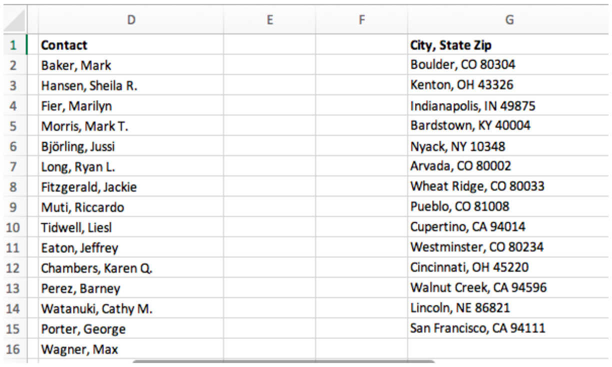
- CONCATENATE (in older Excel versions)
- TEXTJOIN (in Excel versions from 2019, Web — joins text using a text delimiter)

```
=TEXTJOIN(” ”,TRUE, ”The”, ”sun”, ”will”, ”come”,
           ”up”, ”tomorrow.”)
=TEXTJOIN(”, ”, TRUE, A2:A8)
```

Parsing Text Using Text to Columns Feature

- Some columns have complex text that needs to be parsed.
- Excel can parse the text of a column and split it into several columns.
- It’s a bit like when you import a text file.

CH-05.xlsx; watch the video *https://www.linkedin.com/learning/excel-2016-cleaning-up-your-data/split-data-into-columns-with-the-text-to-columns-feature*

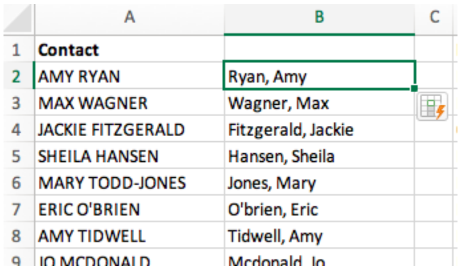


	D	E	F	G
1	Contact			City, State Zip
2	Baker, Mark			Boulder, CO 80304
3	Hansen, Sheila R.			Kenton, OH 43326
4	Fier, Marilyn			Indianapolis, IN 49875
5	Morris, Mark T.			Bardstown, KY 40004
6	Björling, Jussi			Nyack, NY 10348
7	Long, Ryan L.			Arvada, CO 80002
8	Fitzgerald, Jackie			Wheat Ridge, CO 80033
9	Muti, Riccardo			Pueblo, CO 81008
10	Tidwell, Liesl			Cupertino, CA 94014
11	Eaton, Jeffrey			Westminster, CO 80234
12	Chambers, Karen Q.			Cincinnati, OH 45220
13	Perez, Barney			Walnut Creek, CA 94596
14	Watanuki, Cathy M.			Lincoln, NE 86821
15	Porter, George			San Francisco, CA 94111
16	Wagner, Max			

The Magic of Flash Fill

- Flash Fill is one of Excel’s most powerful and least known features.
- Uses AI techniques to try to predict how you want to parse the text.
- Looks like magic, and sometimes might not work for your task.

CH-05.xlsx; watch the video *https://www.linkedin.com/learning/excel-2016-cleaning-up-your-data/use-flash-fill-for-faster-combining-and-splitting*



	A	B	C
1	Contact		
2	AMY RYAN	Ryan, Amy	
3	MAX WAGNER	Wagner, Max	
4	JACKIE FITZGERALD	Fitzgerald, Jackie	
5	SHEILA HANSEN	Hansen, Sheila	
6	MARY TODD-JONES	Jones, Mary	
7	ERIC O'BRIEN	O'brien, Eric	
8	AMY TIDWELL	Tidwell, Amy	
9	JO MCDONALD	Mcdonald, Jo	

2 Cleaning Data in MATLAB

MATLAB's text functions

- *https://au.mathworks.com/help/matlab/characters-and-strings.html*
- MATLAB has two main types for text:
  - Character array — in older versions of MATLAB.
  - String array — more powerful, available since 2016.

Character or string?

- Current versions of MATLAB provide both types but you normally want to work with string arrays only.
- To find out the type of each column (among other things), use summary.

```
>> mlb = readtable('mlb_players.csv');
>> summary(mlb)
Variables:

Name: 1035x1 cell array of character vectors

    Properties:
        Description:  Name
Team: 1035x1 cell array of character vectors

    Properties:
        Description:  Team
Position: 1035x1 cell array of character vectors

    Properties:
        Description:  Position
Height_inches_: 1035x1 double

    Properties:
        Description:  Height(inches)
    Values:

        Min              67
        Median           74
        Max              83
        NumMissing       1

Weight_lbs_: 1035x1 double

    Properties:
        Description:  Weight(lbs)
    Values:

        Min             150
```

```
            Median           200
            Max              290
            NumMissing       2

    Age: 1035x1 double

        Properties:
            Description:   Age
        Values:

            Min              20.9
            Median           27.925
            Max              48.52
            NumMissing       1
```

Setting the Type in a Table Column

- A common problem with MATLAB (and Excel) is that the default settings when reading a CSV file might not be correct.
  - For example, readtable by default may store text as a character array, not a string array.
- If we use MATLAB's import tool we can specify the data type (see lecture week 6).
  - Check how the generated script defines options to the readtable function.
- We can also change the data type *after* the table has been created.

```
mlb.Team = categorical(mlb.Team);
mlb.Name = string(mlb.Name);
```

Filtering Data in an Array

- MATLAB can identify what values meet a particular condition.
- For example, to find what elements in an array ”ages” are larger than 10:

```
>> ages = [1 2 5 34 2 32];
>> ages > 10
ans =
  1x6 logical array
  0   1   0   0   1   0
```

- The result is a filter represented as a *(*logical array): each element is either 0 (”false”) or 1 (”true”).
- We can now select all elements whose corresponding logical array indicates true.

```
>> ages(ages > 10)
ans =
  34   32
```

Filtering Data in a Table

- The same process can be used to remove rows that have columns with some criteria.

```
>> trees.Girth_in_ > 15
ans =
  31x1 logical array
0 0 0 ... 1 1 1
>> wide_trees = trees(trees.Girth_in_ > 15, :)
```

- We can combine multiple filters by using Boolean operators.
- Can you tell what’s the output of the following?

```
>> trees = readtable(”trees.csv”);
>> filtera = trees.Girth_in_ > 10;
>> filterb = trees.Girth_in_ < 15;
>> filterc = trees.Height_ft_ > 70;
>> result = trees(filtera & filterb | filterc , :)
```

Working with Missing Data

- Sometimes, data in some cells are missing.
- In MATLAB, these are indicated with:

  NaN in numerical data.

  NaT in date-time data.

  undefined in categorical data.

- You can do several things with rows that contain NaN.
  1. Ignore the missing data and carry on (almost) as normal.
  2. *Data filtering*: Remove rows with NaN values.
  3. *Data imputation*: Replace NaN cells with guessed values (this is the topic for another unit).

Using Columns with Missing Data

(`https://matlabacademy.mathworks.com/R2020a/portal.html?course=mlvi#chapter=2&lesson=2`)

- Most MATLAB functions generate NaN if one of its inputs contains NaN.

```
avgWS = mean(data.Windspeed)
```

- Some MATLAB functions have an option that allow us to operate with columns that contain missing values.

```
avgP = mean(data.Pressure, ”omitnan”)
```

Removing Rows with Missing Data

- If few rows have columns with missing data we can remove them.
- MATLAB's rmmissing function can remove these rows.
- Sometimes, missing data are represented with unconventional terms, e.g. the string ”N/A”.
- MATLAB's standardizeMissing can mark these as missing data.

```
data = standardizeMissing(data, 'N/A')
data = rmmissing(data)
```

Take-home Messages

Excel

- Fixing problems from manual data input.
- Importing text.
- Text to columns feature.
- Flash Fill.

MATLAB

- Changing data types.
- Text functions.
- Filtering data.
- Removing missing data.

What's Next

- Friday 11 September: Communicator hurdle
- *Mid-semester break 14-25 September*
- Week 8 lecture: Transforming Data