FOSE1025 — Scientific Computing

Week 7 Lecture 1: Cleaning Data

Diego Mollá

FOSE1025 2021H1

Abstract

In this lecture we will focus in the step of data cleaning, with particular emphasis on text data. We will look at various tools that both Excel and MATLAB provide to help cleaning raw data and process text: convert types, parse text, split text, filter data.

Update April 20, 2021

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.

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

These are only some of the functions that can work with text. At the lecture, we will see some of them at work using the file CH-05.xlsx.

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)
```

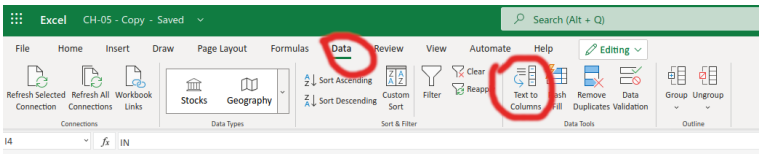
- CONCATENATE (in older Excel versions)
- TEXTJOIN (in Excel versions from 2019, Web — joins text using a text delimiter)

```
=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.
- On Excel Online, look at the "Data" tab

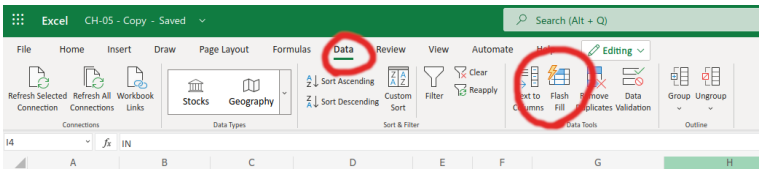
CH-05.xlsx; watch the video *https://www.linkedin.com/learning/excel-2016-cleaning-up-your-data/split-data-into-columns-with-the-text-to-columns-feature*



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*



2 Cleaning Data in MATLAB

MATLAB's Column Types
https://au.mathworks.com/help/matlab/data-types.html

- All values of a MATLAB table column are of the same type.
- Common types in MATLAB are:

  Numeric

  - `double` — This is the default numerical type. It is what is called *double-precision floating point*.
  - There are other types that you can use to represent integers (with or without sign) and other numerical types.

  Text

  - `string` — Starting in MATLAB's version R2016b, this is the preferred way to store text. It's called "string arrays".
  - `char` — Available in all MATLAB versions but not recommended from MATLAB version R2016b. It's called "character arrays".

  Dates and Time

  - `datetime` — MATLAB stores dates and times using the same format.
  - We will look at MATLAB's dates and times in a subsequent lecture.

  Categorical

  - Use this type (instead of, say, `string`, if you know that the column has a finite set of possible values.)
  - For example, `C = categorical({'R','G','B','B','G','B'})` creates a categorical array with six elements that belong to the categories `R`, `G`, or `B`.

Examining the Type of a Table Column
File: biostats.csv

- MATLAB's `summary` function gives a summary of a table.
- It reports various information, including the types of all of its columns.

Try This

1. Use (or generate) the live script that imports the file `biostats.csv` and stores the generated table in the variable `biostats`.
2. Add this command to the live script (without ";" at the end):

```
summary(biostats)
```

The output should look like this:

```
Variables:

Name: 18x1 string

Sex: 18x1 categorical

    Values:

        F         7
        M        11

Age: 18x1 double

    Values:

        Min           23
        Median      32.5
        Max           53

Heightin: 18x1 double

    Values:

        Min           62
        Median      69.5
        Max           75

Weightlbs: 18x1 double

    Values:

        Min           98
        Median       150
        Max          176
```

In this output you can see the type of each column. For columns with categorical data, it will list the number of values in each category. And for columns with numerical data, it will show the minimum, median, and maximum value.

Setting the Type in a Table Column

File: `mlb_players.csv`

- A common problem with MATLAB (and Excel) is that the default settings when reading a CSV file might not be correct.
  - For example, by default, `readtable` 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);
```

In the example `mlb.Team = categorical(mlb.Team);`:

- `mlb.Team` indicates the column with name `Team` which is stored in the table with name `mlb`.
- `categorical(mlb.Team)` returns a column vector where the type of the elements is categorical.
- `mlb.Team = ...` means that the `Team` column of the table `mlb` is assigned the result on the right-hand side of the `=` (which, in our case, is the contents of the same column that has been converted to the categorical type).

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   0   0   1   0   1
```

- 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

File trees.csv

- The same process can be used to select rows whose columns fit 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, :)
```

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.

What's Next

- *Test 2 during this week's SGTA 1*
- Friday 23 April: Communicator hurdle
- Week 8 lecture: Transforming Data