FOSE1025 — Scientific Computing

Week 6 Lecture 1: Cleaning Data

Diego Mollá

FOSE1025 2020H1

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.

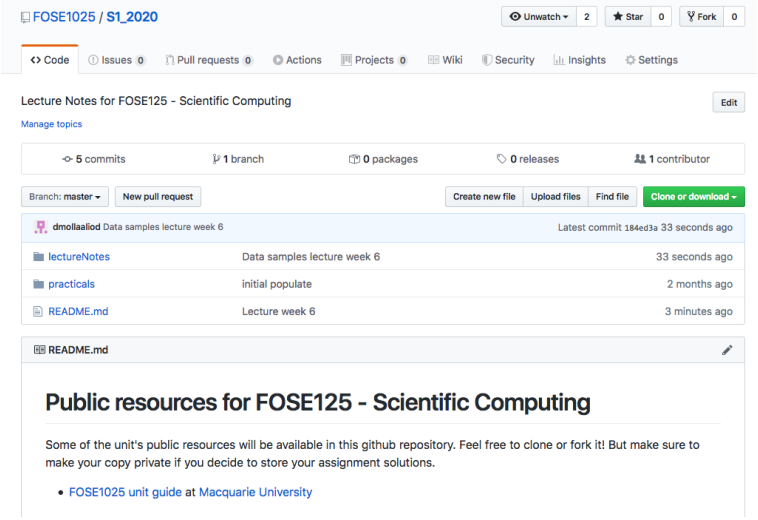
Update April 8, 2020

Contents

Reading

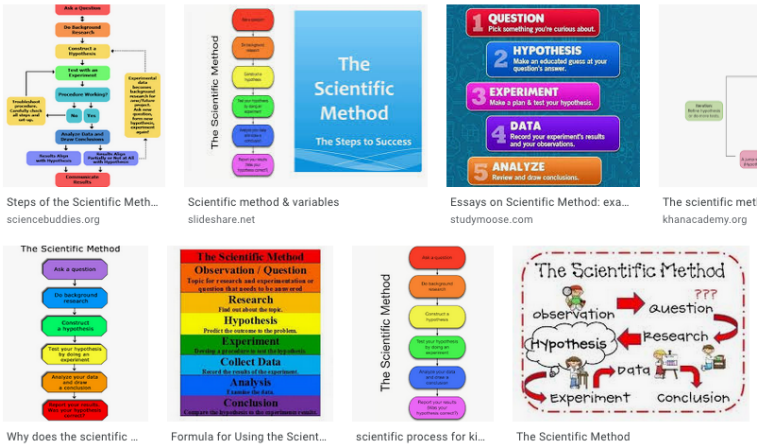
- LinkedIn Learning — Excel 2016: Cleaning up Your Data
  https://www.linkedin.com/learning/excel-2016-cleaning-up-your-data

FOSE1025's public github page



1 Review: Excel for Science

The Scientific Method



Some results of a Google image search with the words "scientific" and "method" — 1 April 2020.

Excel to Manage Data in Science

We are covering these aspects in FOSE1025:

- Import data from external files (e.g. CSV) — Week 3.
- Explore the data — Week 4.
- *Clean the data — Week 6.*
- Preprocess, transform the data — Week 7.
- Analyse, summarise, interpret the data — Week 8.

Importing Data

CSV — Comma Separated Values

- In practice, the file could use other delimiters: tab, semicolon (;), blank space, ...
- Some times, the data fields are determined by the width.

Data Types

- Numbers
- Text
- Dates
- Currency
- ...

Example CSV File

(The lecturer will demo how to import this)

biostats.csv from *https://people.sc.fsu.edu/ jburkardt/data/csv/csv.html*

```
"Name",    "Sex", "Age", "Height (in)", "Weight (lbs)"
"Alex",      "M",    41,          74,         170
"Bert",      "M",    42,          68,         166
"Carl",      "M",    32,          70,         155
"Dave",      "M",    39,          72,         167
"Elly",      "F",    30,          66,         124
"Fran",      "F",    33,          66,         115
"Gwen",      "F",    26,          64,         121
"Hank",      "M",    30,          71,         158
"Ivan",      "M",    53,          72,         175
"Jake",      "M",    32,          69,         143
"Kate",      "F",    47,          69,         139
"Luke",      "M",    34,          72,         163
"Myra",      "F",    23,          62,          98
"Neil",      "M",    36,          75,         160
"Omar",      "M",    38,          70,         145
"Page",      "F",    31,          67,         135
"Quin",      "M",    29,          71,         176
"Ruth",      "F",    28,          65,         131
```

Tables in Excel

(The lecturer will demo how to create and manipulate Excel tables)

- Tables are the fundamental data structure.
- Each row indicates a data sample.
- Each column indicates a type of data.
  - Number, string, date, etc.

– Categorical data: when there is a pre-determined set of values.

	A	B	C	D	E
1	Name	"Sex"	"Age"	"Height (in)"	"Weight (lbs)"
2	Alex	"M"	41	74	170
3	Bert	"M"	42	68	166
4	Carl	"M"	32	70	155
5	Dave	"M"	39	72	167
6	Elly	"F"	30	66	124
7	Fran	"F"	33	66	115
8	Gwen	"F"	26	64	121
9	Hank	"M"	30	71	158
10	Ivan	"M"	53	72	175
11	Jake	"M"	32	69	143
12	Kate	"F"	47	69	139
13	Luke	"M"	34	72	163
14	Myra	"F"	23	62	98
15	Neil	"M"	36	75	160
16	Omar	"M"	38	70	145
17	Page	"F"	31	67	135
18	Quin	"M"	29	71	176
19	Ruth	"F"	28	65	131
20					

Question: What are the data types of each column?

1.1 Data Exploration

2 Cleaning Data

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.*
- But Excel can help find structure from text.

Some Useful Text Functions

CH-06.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
CONCAT	Joins several text items into a cell
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

The Peril of Manual Data Input

- Manual input creates spelling errors.
- Excel has spell checking tools but it is not always useful.
- Spelling errors can be problematic with categorical data.
  - Can be detected by sorting and exploring.

Fixing spelling errors

Option 1

1. Sort by categorical column.
2. Explore and fix.
3. Re-sort by original criteria.

Option 2

1. Apply filter.
2. Explore and fix.
3. Remove filter.

Example: Correct Multiple Misspellings
(The lecturer will demo how to correct spelling mistakes in this table)

CH-06.xlsx; watch the video *https://www.linkedin.com/learning/excel-2016-cleaning-up-your-data/correct-multiple-misspellings*

Employee Name	Building	Department	Benefits	Salary	Job Rating
Baker, Barney	Taft	Executive Education		68,565	3
Barton, Barry	North	Enviromental Health/Safety		91,140	2
Trevino, Gary	South	Professional Training Group	DMR	71,828	1
King, Marye	West	Operations	R	57,158	2
Adkins, Michael	North	Quality Assurance	DM	68,625	5
Fisher, Maria	North	Quality Assurance		74,295	4
Knox, Lori	North	Quality Assurance	DMR	110,340	3
Allison, Timothy	Main	Operations		71,280	1
Rios, Fredrick	North	Enviromental Health/Safety	DMR	86,910	3
Maynard, Susan	South	Executive Education		119,070	1
Bullock, Greg	North	Manufacturing		70,020	4
Ellis, Brenda	West	Peptide Chemistry		96,645	4
Castro, Christopher	Main	Engineering/Maintenance		44,136	5
George, Jessica	North	Process Development	M	69,540	5
Rodgers, Daniel	Mane	Manufacturing	DMR	107,730	2

Parsing Text Using Text to Columns Feature
(The lecturer will demo how to use the "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.

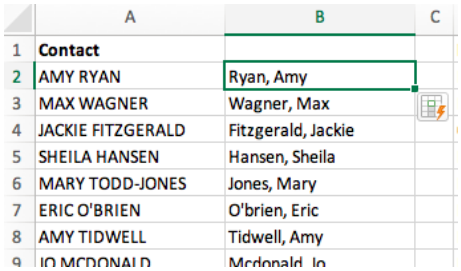
CH-05.xlsx; watch the video *https://www.linkedin.com/learning/excel-2016-cleaning-up-your-data/split-data-into-columns-with-the-text-to-columns-feature*

	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	

Take-home Messages

- Excel as a tool to manage data in science.
- Excel tables.
- Fixing problems from manual data input.
- Importing text.
- Text to columns feature.
- Flash Fill.

What's Next

- Week 7 lecture: Transforming Data
- Week 7, Friday 24 April: Communicator Hurdle