FOSE1025 — Scientific Computing

Week 7 Lecture 1: Transforming Data

Diego Mollá

FOSE1025 2020H1

Abstract

This lecture will focus on the use of Excel for the stage of transforming data for data science projects. The first part will focus on various ways to manipulate times and dates in Excel. We will then see why we might want to normalise data and how we can do it. We will also look at two fundamental ways to represent tables of data: the long format, and the wide format. Finally, we will also introduce pivot tables, which are powerful tools for data transformation and summarisation.

Update April 21, 2020

Contents

Reading

- These notes

1 Dates

This section really belongs to "cleaning data" but we're adding it to this lecture because of time costraints . . . there was enough covered last week already!

Processing Dates

- Dates come in many formats, we need to make sure they are in the format we need.
  - dd/mm/yyyy (Australia)
  - dd.mm.yyyy (Germany)
  - mm/dd/yyyy (USA)
  - yyyy/mm/dd (Japan)
  - . . .

- If input manually, check if there are errors!
  - 24 Maye 2020

Dates Are Not Text or Numbers

- A common problem: treat dates as text or as Numbers. But they are not! They are called “serial numbers” and they represent the number of days since a specific date: 1st January 1900.
- *Exercise 1:* Look at a cell with the date 12/9/22 and change the cell format to text, or change it to number. What do you see?
- *Exercise 2:* Type the number 44818.231 in a cell and change the format to date, what do you see? Change the format now to time. What do you see? What does it all mean??

Useful Functions to Manipulate Dates

Creating Dates and Times

DATE(year,month,day): Create a date from numbers.

TIME(hours,minutes,seconds): Create a time from numbers.

DATE(year,month,day) + TIME(hours,minutes, seconds): Create a date with time.

Useful Functions to Manipulate Dates

Formatting Dates to Text

TEXT(serial_number,pattern)
Represent a date as text using a specific pattern. For example, if cell A1 has the date =DATE(2020,12,23) + TIME(21,35,12):

TEXT(A1, ”dd/mm/yy”) has the value ”23/12/20”

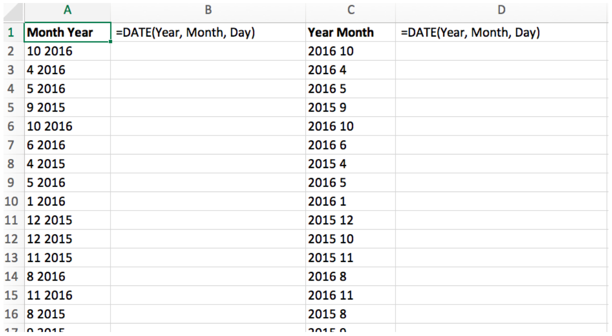
TEXT(A1, ”dd/mm/yyyy hh:mm”) has the value ”23/12/2020 21:35”

TEXT(A1, ”dd mmm yyyy hh:mm:ss”) has the value ”23 Dec 2020 21:35:12” (notice the three “m”?)

TEXT(A1, ”dd mmmm yyyy hh:mm AM/PM”) has the value ”23 December 2020 09:35 pm”

Example 1: Dates in Different Formats

Ch-03.xlsx from *https://www.linkedin.com/learning/excel-2016-cleaning-up-your-data*



	A	B	C	D
1	**Month Year**	=DATE(Year, Month, Day)	**Year Month**	=DATE(Year, Month, Day)
2	10 2016		2016 10	
3	4 2016		2016 4	
4	5 2016		2016 5	
5	9 2015		2015 9	
6	10 2016		2016 10	
7	6 2016		2016 6	
8	4 2015		2015 4	
9	5 2016		2016 5	
10	1 2016		2016 1	
11	12 2015		2015 12	
12	12 2015		2015 10	
13	11 2015		2015 11	
14	8 2016		2016 8	
15	11 2016		2016 11	
16	8 2015		2015 8	
17	9 2015		2015 9	

Exercise: Mixed date formats in one column

Create a blank Excel worksheet, import this CSV file, and normalise the dates.

dates.csv

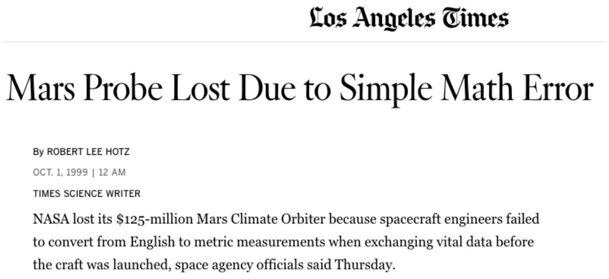
```
Date,Name,Email,Consultation Times,Zoom
12/01/2020,Diego Molla-Aliod,diego.molla-aliodmq.edu.au,,12 May 2020,Gaurav Gupta,gaurav.guptamq.edu
15 Apil 2020,Urvashi Khanna,urvashi.khannamq.edu.au,Wed 12-1,https://macquarie.zoom.us/j/4725684612020-
11-23,Munazza Zaib,munazza-zaibmq.edu.au,Wed 11-12,https://macquarie.zoom.us/j/267542550
```

- If you just double-click on the CSV file and let Excel import the file using defaults, the resulting dates look strange... why?
- Hint: don't let Excel use the General format for the first column.

2 Measures

Normalising Measures

- Beware with the measures.
- One column might be in metres, another in centimetres, another in inches ...



Los Angeles Times

Mars Probe Lost Due to Simple Math Error

By ROBERT LEE HOTZ

OCT. 1, 1999 | 12 AM

TIMES SCIENCE WRITER

NASA lost its $125-million Mars Climate Orbiter because spacecraft engineers failed to convert from English to metric measurements when exchanging vital data before the craft was launched, space agency officials said Thursday.

Re-scaling

- Depending on the application (e.g. for machine learning), one may want to ensure that all values are within a certain range.
- For example, between 0 and 1:
  1. Identify the minimum and the maximum of all values
  2. Change all values using this formula:

$$newvalue = \frac{oldvalue - minimum}{maximum - minimum}$$

- Another common normalisation approach uses the *mean* and the *standard deviation.*
  1. Calculate the mean (in Excel: AVERAGE) and the standard deviation (in Excel: STDEV.P)
  2. Change all values using this formula:

$$newvalue = \frac{oldvalue - mean}{stdev}$$

Example: Normalising Values in Excel

Can you normalise the girth, height, and volume?

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

```
 "Index", "Girth (in)", "Height (ft)", "Volume(ft^3)"
 1,   8.3,     70,   10.3
 2,   8.6,     65,   10.3
 3,   8.8,     63,   10.2
 4,  10.5,     72,   16.4
 5,  10.7,     81,   18.8
 6,  10.8,     83,   19.7
 7,  11.0,     66,   15.6
 8,  11.0,     75,   18.2
 9,  11.1,     80,   22.6
10,  11.2,     75,   19.9
11,  11.3,     79,   24.2
12,  11.4,     76,   21.0
13,  11.4,     76,   21.4
14,  11.7,     69,   21.3
15,  12.0,     75,   19.1
16,  12.9,     74,   22.2
17,  12.9,     85,   33.8
18,  13.3,     86,   27.4
19,  13.7,     71,   25.7
20,  13.8,     64,   24.9
21,  14.0,     78,   34.5
22,  14.2,     80,   31.7
23,  14.5,     74,   36.3
24,  16.0,     72,   38.3
25,  16.3,     77,   42.6
26,  17.3,     81,   55.4
27,  17.5,     82,   55.7
28,  17.9,     80,   58.3
29,  18.0,     80,   51.5
30,  18.0,     80,   51.0
31,  20.6,     87,   77.0
```

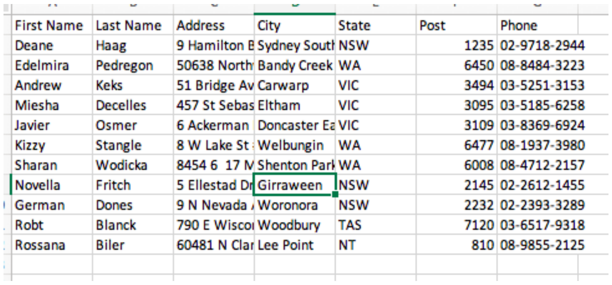
3 Long and Wide Formats

3.1 Long and Wide Formats

Tables as 2D Data

- Remember that tables represent 2-dimensional information.
  - Rows indicate different records.
  - Columns indicate different types of data in the record.
- We can, for example, represent the work address (street, city, postcode, etc) of a group of people.

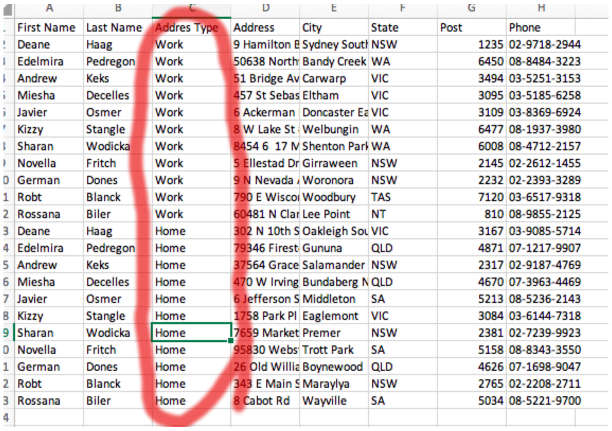
(file WorkAddresses.xlsx)



First Name	Last Name	Address	City	State	Post	Phone
Deane	Haag	9 Hamilton B	Sydney South	NSW	1235	02-9718-2944
Edelmira	Pedregon	50638 North	Bandy Creek	WA	6450	08-8484-3223
Andrew	Keks	51 Bridge Av	Carwarp	VIC	3494	03-5251-3153
Miesha	Decelles	457 St Sebas	Eltham	VIC	3095	03-5185-6258
Javier	Osmer	6 Ackerman	Doncaster Ea	VIC	3109	03-8369-6924
Kizzy	Stangle	8 W Lake St	Welbungin	WA	6477	08-1937-3980
Sharan	Wodicka	8454 6 17 M	Shenton Park	WA	6008	08-4712-2157
Novella	Fritch	5 Ellestad Dr	Girraween	NSW	2145	02-2612-1455
German	Dones	9 N Nevada /	Woronora	NSW	2232	02-2393-3289
Robt	Blanck	790 E Wiscor	Woodbury	TAS	7120	03-6517-9318
Rossana	Biler	60481 N Clar	Lee Point	NT	810	08-9855-2125

Tables as 3D, 4D . . . ?

- How would you keep information about the work *and the home address*?
- What if one person has 15 different properties, how do you store the information for all people?
- A solution: Add one column that indicates the type of address.
- (Databases can encode this information more efficiently using relational tables but this is not the topic of this unit.)

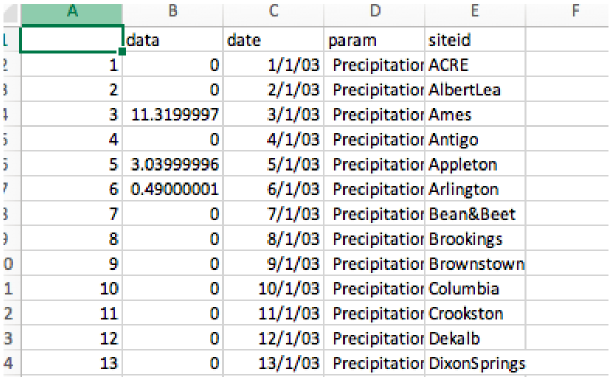


First Name	Last Name	Addres Type	Address	City	State	Post	Phone
Deane	Haag	Work	9 Hamilton B	Sydney South	NSW	1235	02-9718-2944
Edelmira	Pedregon	Work	50638 North	Bandy Creek	WA	6450	08-8484-3223
Andrew	Keks	Work	51 Bridge Av	Carwarp	VIC	3494	03-5251-3153
Miesha	Decelles	Work	457 St Sebas	Eltham	VIC	3095	03-5185-6258
Javier	Osmer	Work	6 Ackerman	Doncaster Ea	VIC	3109	03-8369-6924
Kizzy	Stangle	Work	8 W Lake St	Welbungin	WA	6477	08-1937-3980
Sharan	Wodicka	Work	8454 6 17 M	Shenton Park	WA	6008	08-4712-2157
Novella	Fritch	Work	5 Ellestad Dr	Girraween	NSW	2145	02-2612-1455
German	Dones	Work	9 N Nevada /	Woronora	NSW	2232	02-2393-3289
Robt	Blanck	Work	790 E Wiscor	Woodbury	TAS	7120	03-6517-9318
Rossana	Biler	Work	60481 N Clar	Lee Point	NT	810	08-9855-2125
Deane	Haag	Home	302 N 10th S	Oakleigh Sou	VIC	3167	03-9085-5714
Edelmira	Pedregon	Home	79346 Firest	Gununa	QLD	4871	07-1217-9907
Andrew	Keks	Home	37564 Grace	Salamander	NSW	2317	02-9187-4769
Miesha	Decelles	Home	470 W Irving	Bundaberg N	QLD	4670	07-3963-4469
Javier	Osmer	Home	6 Jefferson S	Middleton	SA	5213	08-5236-2143
Kizzy	Stangle	Home	1758 Park Pl	Eaglemont	VIC	3084	03-6144-7318
Sharan	Wodicka	Home	7659 Market	Premer	NSW	2381	02-7239-9923
Novella	Fritch	Home	95830 Webs	Trott Park	SA	5158	08-8343-3550
German	Dones	Home	26 Old Willia	Boynewood	QLD	4626	07-1698-9047
Robt	Blanck	Home	343 E Main S	Maraylya	NSW	2765	02-2208-2711
Rossana	Biler	Home	8 Cabot Rd	Wayville	SA	5034	08-5221-9700

Long and Wide Formats

- The tables that we are used to see are in the *wide format*.
  - Each column indicates a specific data: name, address, location, temperature, etc.
- For complex data we may want to use a *long format*.
  - One column indicates the type of data.
  - Another column (or columns) indicate the value.

(file weather_data.csv)



A	B	C	D	E	F
	data	date	param	siteid	
1	0	1/1/03	Precipitatior	ACRE	
2	0	2/1/03	Precipitatior	AlbertLea	
3	11.3199997	3/1/03	Precipitatior	Ames	
4	0	4/1/03	Precipitatior	Antigo	
5	3.03999996	5/1/03	Precipitatior	Appleton	
6	0.49000001	6/1/03	Precipitatior	Arlington	
7	0	7/1/03	Precipitatior	Bean&Beet	
8	0	8/1/03	Precipitatior	Brookings	
9	0	9/1/03	Precipitatior	Brownstown	
10	0	10/1/03	Precipitatior	Columbia	
11	0	11/1/03	Precipitatior	Crookston	
12	0	12/1/03	Precipitatior	Dekalb	
13	0	13/1/03	Precipitatior	DixonSprings	

Processing Tables in Long Format

The lecturer will demonstrate how to use filters and pivot tables to process tables in long format

- Many tables are expressed in long format for some columns.
- Excel does not have a specific tool to process these tables.
- You can use filters to focus on specific values.
- You can also use *pivot tables*.
- We will see pivot tables more in detail next week, but here we see how to use them to process tables in long format.

3.2 Introducing Pivot Tables

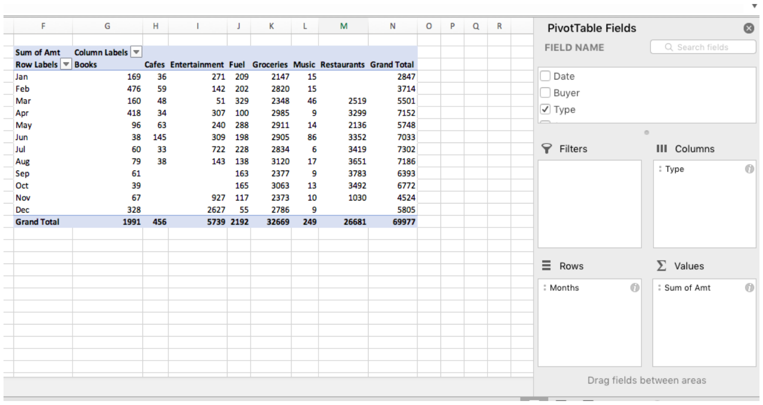
This section really belongs to next week's data summarisation, we will see more of this, and data analysis, next week.

Pivot Tables: A Motivational Example

(data from *https://www.linkedin.com/learning/excel-pivottables-for-beginners*)

- Find the total shopping in each category "Fuel", etc, of file shopping.csv.
- Find the total shopping of each month.
- What shopping per month and per category??
- Pivot tables can help you generate data for all of above and more.

A Simple Pivot Table



Sum of Amt	Column Labels							
Row Labels	Books	Cafes	Entertainment	Fuel	Groceries	Music	Restaurants	Grand Total
Jan	169	36	271	209	2147	15		2847
Feb	476	59	142	202	2820	15		3714
Mar	160	48	51	329	2348	46	2519	5501
Apr	418	34	307	100	2985	9	3299	7152
May	96	63	240	288	2911	14	2136	5748
Jun	38	145	309	198	2905	86	3352	7033
Jul	60	33	722	228	2834	6	3419	7302
Aug	79	38	143	138	3120	17	3651	7186
Sep	61			163	2377	9	3783	6393
Oct	39			165	3063	13	3492	6772
Nov	67		927	117	2373	10	1030	4524
Dec	328		2627	55	2786	9		5805
Grand Total	1991	456	5739	2192	32669	249	26681	69977

PivotTable Fields

FIELD NAME

□ Date
□ Buyer
☑ Type

Filters

Columns: Type

Rows: Months

Values: Sum of Amt

Drag fields between areas

Anatomy of a Pivot Table

Filters

- What column to use to filter values.
- Only for columns with categorical data.

Rows

- What column to use in the rows of the pivot table.
- Only for columns with categorical data.

Columns

- What column to use in the columns of the pivot table.
- Only for columns with categorical data.

Values

- What value we want to aggregate.
- Only for columns with numerical data.

Take-home Messages

- Dates and times are all the same thing in Excel . . . and very different to other data types!
- Pay attention when importing files that use unconventional date and time expressions.
- You need to understand why you want to normalise data, and be able to do some simple data normalisation.
- You need to be explain the difference between wide and long formats, and process tables in each kind of format.

What's Next

- Week 8 lecture: Summarising, Visualising and Analysing Data.
- Week 8: in-class quiz before the lecture.