FOSE1025 — Scientific Computing

Week 8 Lecture 1: Transforming Data

Diego Mollá

FOSE1025 2023H1

Abstract

This lecture will focus on the stage of transforming data for data science projects. The first part will focus on various ways to manipulate times and dates in Excel and MATLAB. We will then look at two fundamental ways to represent tables of data: the long format, and the wide format. Finally, we will introduce Excel's pivot tables, which are powerful tools for data transformation and summarisation.

Update April 18, 2023

Contents

Reading

- These notes
- Readings listed in iLearn — Week 8

1 Dates

This section really belongs to "cleaning data" but we're adding it to this lecture because of time constraints . . . 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

What might we want to do with dates?

- Sort by date
- Group (or plot) by month, by year, etc
- Determine the time elapsed between two dates or times

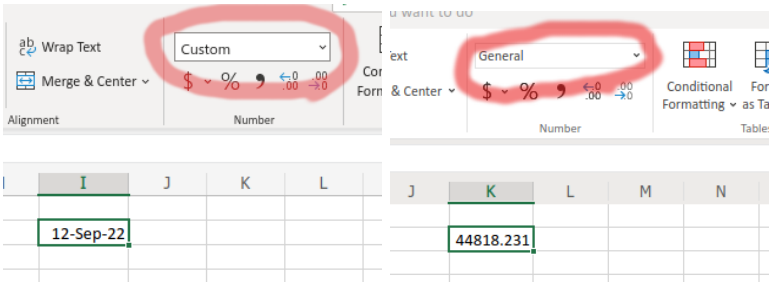
1.1 Dates in Excel

Excel Dates and Times Are Numbers

Internally, Excel stores dates and times as numbers. These are called "serial numbers" and they represent the number of days since a specific date: 1st January 1900.

Demonstration 1
Type `12 Sep 2022` in an Excel cell and observe how it shows the date (see screenshot). Change the cell format to "Number". You will see that the cell now displays the number `44816.00`.

Demonstration 2
Type the number `44818.231` in an Excel cell and change the format to "Short Date". You will see the date `9/14/2022`. Change the format now to "Time." You will see the time `5:32:38 AM`.



The second demonstration shows that the serial number contains the information of *both the date and the time*:

- The integer part of the number indicates the number of days since 1st January 1900. In our example, the date 14 Sep 2022 is the day that happens 44818 days after 1 Jan 1900. Incidentally, note that the date is displaying using the US convention: month/day/year. This is the default in my computer, it may display differently in your computer.
- The fractional part of the number indicates the time as the fraction of day. For example, take the number 0.231. If you want to know the number of hours that corresponds to the fraction 0.231, multiply the number by 24 (because a day has 24 hours). The integer part of the result ($0.231 \times 24 = 5.775$) says that the number of hours is 5. Can you now figure out how to obtain the number of minutes (32) and seconds (38)?

Useful Excel Functions to Manipulate Dates and Times
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 Excel Functions to Manipulate Dates and Times
From Dates and Times to Text

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

TEXT(A1, "dd/mm/yy") returns the text "23/12/20".

TEXT(A1, "dd/mm/yyyy hh:mm") returns the text "23/12/2020 21:35".

TEXT(A1, "dd mmm yyyy hh:mm:ss") returns the text "23 Dec 2020 21:35:12" (notice the three "m"?).

TEXT(A1, "dd mmmm yyyy hh:mm AM/PM") returns the text "23 December 2020 09:35 pm" (now there are four "m").

Useful Excel Functions to Manipulate Dates and Times
From Text to Dates and Times

DATEVALUE(text)
Convert text into a serial number that represents the date. This function does not convert times, only dates.

DATEVALUE("12 May 2021") returns the number 44328.

DATEVALUE("12 May 2021 3:15pm") returns the same number 44328.

VALUE(text)
Convert text into a serial number that represents the date and time.

VALUE("12 May 2021") returns the number 44328.

VALUE("12 May 2021 3:15pm") returns the same number 44328.64.

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

1. What formula would you type in cell B2?
2. What formula would you type in cell D2?

	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	

Exercise: Mixed date formats in one column

Create a blank Excel worksheet, import this CSV file, and normalise the dates so that they appear as in the screenshot.

dates.csv

	A	B	C	D	E	F	G	H
1	Date	Name	Email	Consultati	Zoom			
2	Thursday, December 1, 2022	Diego Mol	diego.molla-aliod@mq.edu.au					
3	Thursday, May 12, 2022	Charanya I	charanya.ramakrishnan@mq.edu.au					
4	Friday, April 15, 2022	Urvashi Kh	urvashi.kh	Thu 11am	room 4RPD G02			
5	Wednesday, November 23, 2022	Munazza Z	munazza-z	Fri 21-1pm	https://macquarie.zoom.us/j/85387376629			
6	Friday, May 13, 2022	Sepehr (Se	sepehr.tor	Wed 10-1	room 4RPD G02			
7	Wednesday, November 23, 2022	Hubert Ha	hubert.hai	Thu 1-2pm	https://macquarie.zoom.us/j/83256626172			

Content of dates.csv:

```
Date,Name,Email,Consultation Times,Zoom
12/01/2022,Diego Molla-Aliod,diego.molla-aliod@mq.edu.au,,
12 May 2022,Charanya Ramakrishnan,charanya.ramakrishnan@mq.edu.au,,
15 Apil 2022,Urvashi Khanna,urvashi.khanna@mq.edu.au,Thu 11am-12pm,room 4RPD G02
2022-11-23,Munazza Zaib,munazza-zaib@mq.edu.au,Fri 21-1pm, https://macquarie.zoom.us/j
    /85387376629
05/13/2022,Sepehr (Sep) Torfeh Nejad,sepehr.torfehnejad@mq.edu.au,Wed 10-11am, room 4RPD G02
"Nov 23, 2022",Hubert Hartan,hubert.hartan@mq.edu.au,Thu 1-2pm,  https://macquarie.zoom.us/j
    /83256626172
```

Excel did a good job to guess the date from the input CSV but the resulting worksheet is trying to display them using the original date format. You should have no problems to display the dates using the long date format. The only problem was with cell A4, which has a typo. You can do this:

1. Select column A and set the format to "Long Date".
2. Edit cell A4 to correct the typo in the cell. After correcting the typo, Excel will correctly convert the cell to the date in the correct format.

Operating with Excel Dates

Extraction

The following commands extract parts of a date and time:

- `YEAR`, `MONTH`, `DAY`, `HOUR`, `MINUTE`, `SECOND`
- The result is a *number*, not a date (i.e. not a serial number)

Time difference

- The following command can be used to find the difference between two dates:
  - `DATEDIF(date1, date2, "y")` — difference in years
  - `DATEDIF(date1, date2, "m")` — difference in months
  - `DATEDIF(date1, date2, "d")` — difference in days
  - Again, the result is a number, not a date.
- If cells `A2` and `B2` contain dates, then:
  - `B2 - A2` is the time difference in days (and fraction of days).

In the above examples, we presume that the time shown in cell A2 is before the time shown in B2. Otherwise, the function `DATEDIF` will generate an error.

1.2 Dates in MATLAB

Understanding Dates in MATLAB

https://au.mathworks.com/help/matlab/date-and-time-operations.html

- As with Excel, MATLAB has a specific data format for date-time.
- MATLAB's `datetime` allows one to create a date-time. It accepts several formats, including:
  - year, month, day
  - year, month, day, hour, minute, second

```
hello_date = datetime(2020, 7, 3, 18, 30, 23)
hello_date = datetime(2020, 7, 3)
```

MATLAB does not try to guess the meaning of each number. *They must be placed in the correct order.* Compare these:

```
date1 = datetime(2020, 7, 3)
date2 = datetime(2020, 3, 7)
date3 = datetime(3, 7, 2020)
```

In the above examples, MATLAB will take the first number as the year, the second as the month, and the third as the day (so the third line will generate a completely unexpected date).

From Text to Dates and Back

- Sometimes we want to convert text containing a date (and/or time) into MATLAB's date-time, or vice-versa.
- MATLAB's `datetime` can convert from text (and other types) to date.

```
t = datetime('21/09/2020')
```

This example converts the string '21/09/2020' into a MATLAB date.

- MATLAB's `string` converts from date (and other types) to text.

```
w_table.StringDate = string(w_table.Date,
                            'MM/dd/yyyy')
```

This example converts all dates from the `Date` column of table with name `w_table` into text. The result is then stored in column with name `StringDate` of the same table `w_table`. The format 'MM/dd/yyyy' is used for the conversion to text.

Text Date Formats in MATLAB
https://au.mathworks.com/help/matlab/ref/datetime.html#buhzxmk-1-Format

- MATLAB allows to read and write dates using different formats.
- MATLAB formats are slightly different from Excel's formats.
- These formats can also be used when importing from CSV files.

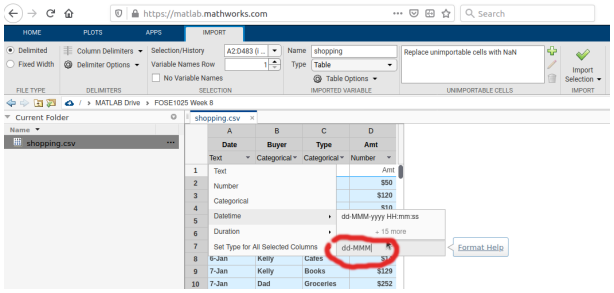
Examples of Formats

Format	Example
dd-MMM-yyyy HH:mm:ss	01-Mar-2000 15:45:17
MM/dd/yyyy	03/01/2000
MM dd yy	03 01 00

Note different MATLAB functions may use different variants of time format specifications. For example, the function "datestr" converts dates to text using a different pattern (*https://au.mathworks.com/help/matlab/ref/datestr.htm*
If in doubt, read the MATLAB documentation.

Example: Importing shopping.csv

- The file shopping.csv represents dates using the day and month only, using a specific format of the form "1-Jan", "2-Jan", etc.
- In MATLAB, specify the datetime format "dd-MMM" in the "Date" column when you use the data import wizard.



2 Long and Wide Formats

2.1 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 E	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 A	Woronora	NSW	2232	02-2393-3289
Robt	Blanck	790 E Wisco	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.)

	A	B	C	D	E	F	G	H
	First Name	Last Name	Addres Type	Address	City	State	Post	Phone
1	Deane	Haag	Work	9 Hamilton E	Sydney South	NSW	1235	02-9718-2944
2	Edelmira	Pedregon	Work	50638 North	Bandy Creek	WA	6450	08-8484-3223
3	Andrew	Keks	Work	51 Bridge Av	Carwarp	VIC	3494	03-5251-3153
4	Miesha	Decelles	Work	457 St Sebas	Eltham	VIC	3095	03-5185-6258
5	Javier	Osmer	Work	6 Ackerman	Doncaster Ea	VIC	3109	03-8369-6924
6	Kizzy	Stangle	Work	8 W Lake St	Welbungin	WA	6477	08-1937-3980
7	Sharan	Wodicka	Work	8454 6 17 M	Shenton Park	WA	6008	08-4712-2157
8	Novella	Fritch	Work	5 Ellestad Dr	Girraween	NSW	2145	02-2612-1455
9	German	Dones	Work	9 N Nevada A	Woronora	NSW	2232	02-2393-3289
10	Robt	Blanck	Work	790 E Wisco	Woodbury	TAS	7120	03-6517-9318
11	Rossana	Biler	Work	60481 N Clar	Lee Point	NT	810	08-9855-2125
12	Deane	Haag	Home	302 N 10th S	Oakleigh Sou	VIC	3167	03-9085-5714
13	Edelmira	Pedregon	Home	79346 Firest	Gununa	QLD	4871	07-1217-9907
14	Andrew	Keks	Home	37564 Grace	Salamander	NSW	2317	02-9187-4769
15	Miesha	Decelles	Home	470 W Irving	Bundaberg N	QLD	4670	07-3963-4469
16	Javier	Osmer	Home	6 Jefferson S	Middleton	SA	5213	08-5236-2143
17	Kizzy	Stangle	Home	1758 Park Pl	Eaglemont	VIC	3084	03-6144-7318
18	Sharan	Wodicka	Home	7659 Market	Premer	NSW	2381	02-7239-9923
19	Novella	Fritch	Home	9583 30 Webs	Trott Park	SA	5158	08-8343-3550
20	German	Dones	Home	26 Old Willia	Boynewood	QLD	4626	07-1698-9047
21	Robt	Blanck	Home	343 E Main S	Maraylya	NSW	2765	02-2208-2711
22	Rossana	Biler	Home	8 Cabot Rd	Wayville	SA	5034	08-5221-9700
23								

Long and Wide Formats

- Most tables that we have seen in this unit so far are in a *wide format*.
  - Each column indicates 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
1		data	date	param	siteid	
2	1	0	1/1/03	Precipitatior	ACRE	
3	2	0	2/1/03	Precipitatior	AlbertLea	
4	3	11.3199997	3/1/03	Precipitatior	Ames	
5	4	0	4/1/03	Precipitatior	Antigo	
6	5	3.03999996	5/1/03	Precipitatior	Appleton	
7	6	0.49000001	6/1/03	Precipitatior	Arlington	
8	7	0	7/1/03	Precipitatior	Bean&Beet	
9	8	0	8/1/03	Precipitatior	Brookings	
10	9	0	9/1/03	Precipitatior	Brownstown	
11	10	0	10/1/03	Precipitatior	Columbia	
12	11	0	11/1/03	Precipitatior	Crookston	
13	12	0	12/1/03	Precipitatior	Dekalb	
14	13	0	13/1/03	Precipitatior	DixonSprings	

In this example, the column with name "param" indicates the type of data, and the column "data" indicates the value. For example, since cell D4 has the value "Precipitation", then cell B4 indicates that the precipitation of site Ames on 3/1/03 was 11.3199997.

Processing Excel Tables in Long Format
The lecturer will demonstrate how to use filters and pivot tables to process Excel tables in long format. File: shopping.csv

- Many tables are expressed in long format for some columns.
- Excel does not have a specific tool to process these tables.
- You have seen how you can use filters to focus on specific values.
- You have also seen how you can use conditional functions to calculate values of one column based on the values of another column.
  - e.g. `=SUMIFS(D:D,C:C,"Fuel")` sums all values in column D such that the cell in column C has the value "Fuel").
- You can also use *pivot tables*.

2.2 Introducing Pivot Tables

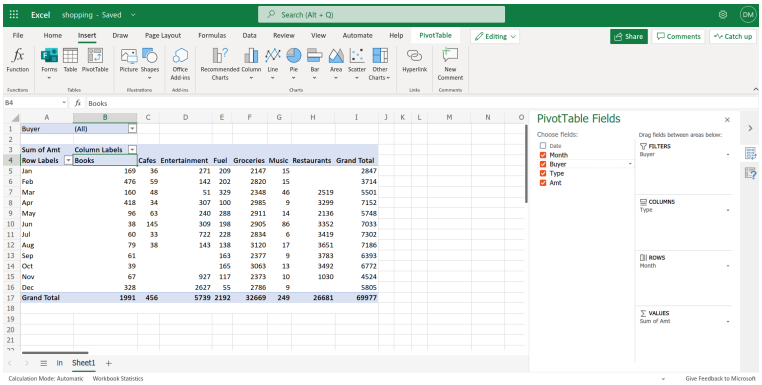
This section really belongs to next week's data summarisation. We will see more of this, and data analysis, next week.

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.

	Date	Buyer	Type	Amt
1	1-Jan	Mom	Fuel	$50
2	2-Jan	Mom	Groceries	$120
3	3-Jan	Dad	Cafes	$10
4	4-Jan	Dad	Fuel	$40
5	4-Jan	Kelly	Groceries	$129
6	5-Jan	Mom	Cafes	$12
7	6-Jan	Kelly	Cafes	$14
8	7-Jan	Kelly	Books	$129
9	7-Jan	Dad	Groceries	$252
10	9-Jan	Kelly	Fuel	$44
11	10-Jan	Dad	Groceries	$39
12	12-Jan	Mom	Books	$20
13	13-Jan	Dad	Groceries	$132
14	14-Jan	Dad	Groceries	$79
15	16-Jan	Kelly	Groceries	$172
16	16-Jan	Dad	Music	$8
17	18-Jan	Kelly	Fuel	$30

A Simple Pivot Table



Buyer	(All)							
Sum of Amt	Column Labels							
Row Labels	Books	Cafes	Entertainment	Fuel	Groceries	Music	Restaurants	Grand Total
Jan	169	96		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

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.

Pivot Tables that Convert from Long to Wide

Exercise 1 (weather_data.csv)
What is the average precipitation in Antigo?

- Using AVERAGEIFS
- Using a pivot table

Exercise 2 (weather_data.csv)
What is the March-2013 average precipitation in Antigo?

- Using AVERAGEIFS
- Using a pivot table

	A	B	C	D	E	F
1		data	date	param	siteid	
2	1	0	1/1/03	Precipitatior	ACRE	
3	2	0	2/1/03	Precipitatior	AlbertLea	
4	3	11.3199997	3/1/03	Precipitatior	Ames	
5	4	0	4/1/03	Precipitatior	Antigo	
6	5	3.03999996	5/1/03	Precipitatior	Appleton	
7	6	0.49000001	6/1/03	Precipitatior	Arlington	
8	7	0	7/1/03	Precipitatior	Bean&Beet	
9	8	0	8/1/03	Precipitatior	Brookings	
10	9	0	9/1/03	Precipitatior	Brownstown	
11	10	0	10/1/03	Precipitatior	Columbia	
12	11	0	11/1/03	Precipitatior	Crookston	
13	12	0	12/1/03	Precipitatior	Dekalb	
14	13	0	13/1/03	Precipitatior	DixonSprings	

Take-home Messages

- Both Excel and MATLAB have a specific data type that is used to represent Dates and times.
- Pay attention when importing files that do not use conventional date and time expressions. In those cases, Excel and MATLAB may guess the wrong format.
- Both Excel and MATLAB offer functions that can be used to create dates and convert dates to text.
- You need to understand the power of Excel's pivot tables.

What's Next

Assessments this week

- Problem Solver employability hurdle: Friday 28 April

Next Week

- Week 9 lecture: Summarising, Visualising and Analysing Data.