FOSE1025 — Scientific Computing

Week 9 Lecture 1: Summarising and Analysing Data

Diego Mollá

FOSE1025 2021H2

Abstract

This lecture will focus on several approaches for summarising and preparing the data for the final analysis. We will look at pivot tables as a powerful tool to transform and summarising the data. With pivot tables we can convert tables from the long to the wide format. In addition, we can aggregate and filter data and make it ready for insightful analysis and graphic representations. Beside pivot tables, we will look at some specific tools that Excel and MATLAB provide for the analysis of data.

Update September 30, 2021

Contents

1 Excel's Pivot Tables **2**
1.1 Excel's Pivot Tables for Charts 4

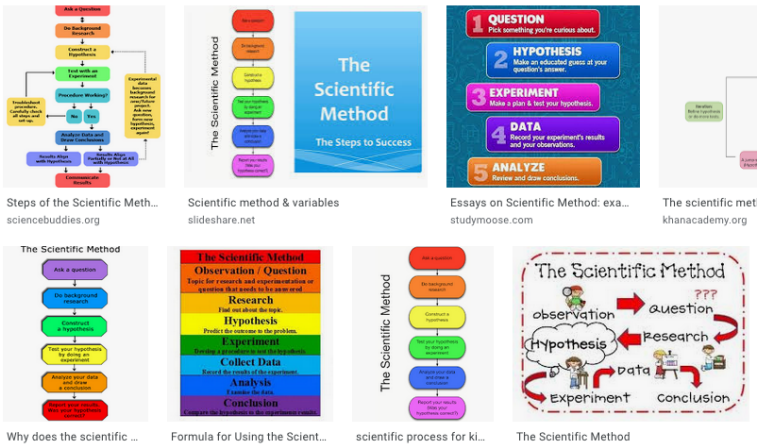
2 Processing Long Tables in MATLAB **5**

3 Data Analysis in MATLAB **8**
3.1 Plotting in MATLAB 8
3.2 MATLAB: Finding Correlations 9

Reading

- These notes
- Related MATLAB scripts
- *https://au.mathworks.com/help/releases/R2020a/matlab/ref/double.groupsummary.html*
- *https://au.mathworks.com/help/releases/R2020a/matlab/ref/unstack.html*

The Scientific Method



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

Excel and MATLAB to Manage Data in Science

We are covering these aspects in FOSE1025:

- Represent data in Excel — Weeks 2 & 3.
- Represent data in MATLAB — Weeks 3 & 5.
- Explore data in Excel — Week 4.
- Visualise data in Excel — Week 5.
- Import data from external files (e.g. CSV) — Week 6.
- MATLAB scripts for reproducibility — Week 6.
- Clean the data (Excel, MATLAB) — Week 7.
- Preprocess, transform the data (Excel, MATLAB) — Week 8.
- *(you are here)*
- Analyse, summarise, interpret the data (MATLAB) — Week 9.
- Ethics of Data — Week 10.

1 Excel's Pivot Tables

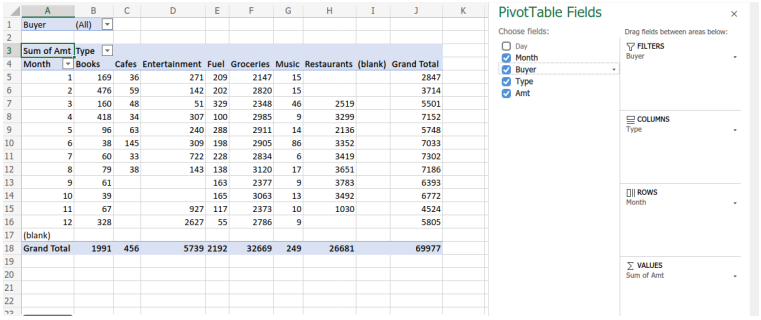
Pivot Tables: A Motivational Example

(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
File: shopping.csv



Buyer	(All)									
Sum of Amt	Type									
Month	Books	Cafes	Entertainment	Fuel	Groceries	Music	Restaurants	(blank)	Grand Total	
1	169	36	271	209	2147	15			2847	
2	476	59	142	202	2820	15			3714	
3	160	48	51	329	2348	46	2519		5501	
4	418	34	307	100	2985	9	3299		7152	
5	96	63	240	288	2911	14	2136		5748	
6	38	145	309	198	2905	86	3352		7033	
7	60	33	722	228	2834	6	3419		7302	
8	79	38	143	138	3120	17	3651		7186	
9	61			163	2377	9	3783		6393	
10	39			165	3063	13	3492		6772	
11	67		927	117	2373	10	1030		4524	
12	328		2627	55	2786	9			5805	
(blank)										
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 to 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?

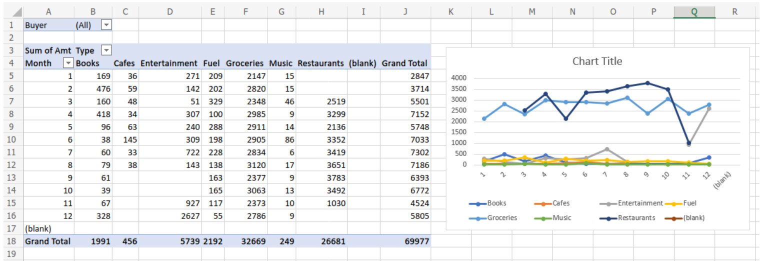
- Using AVERAGEIFS
- Using a pivot table

	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	

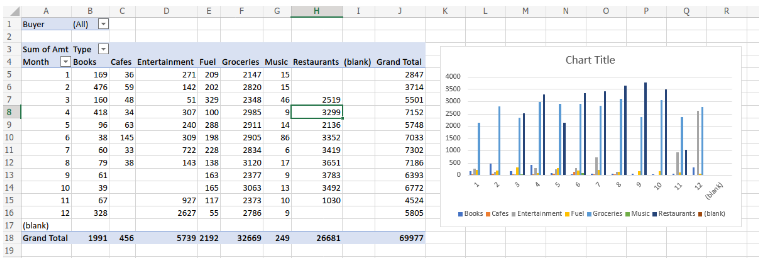
1.1 Excel's Pivot Tables for Charts

Pivot Tables for Charts
Use file shopping.csv

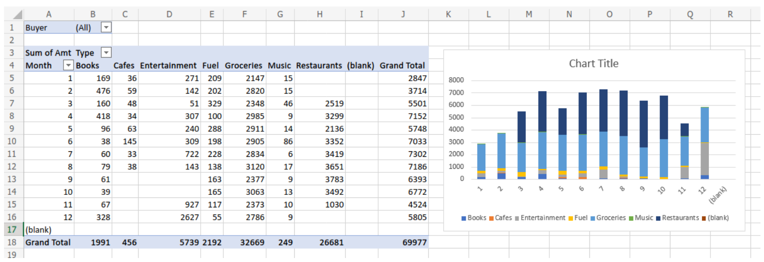
- Pivot tables facilitate the transformation of data for the creation of complex plots.
- In a *multiple chart*, each column of a table is plotted overlayed with the rest. Good for line charts.
- In a *clustered chart*, each row forms a cluster.
- In a *stacked chart*, the data of a table are plotted one on top of the other. Good for column charts.



Sum of Amt Month	Books	Cafes	Entertainment	Fuel	Groceries	Music	Restaurants	(blank)	Grand Total
1	169	36	271	209	2147	15			2847
2	476	59	142	202	2820	15			3714
3	160	48	51	329	2348	46	2519		5501
4	418	34	307	100	2985	9	3299		7152
5	96	63	240	288	2911	14	2136		5748
6	38	145	309	198	2905	86	3352		7033
7	60	33	722	228	2834	6	3419		7302
8	79	38	143	138	3120	17	3651		7186
9	61			163	2377	9	3783		6393
10	39			165	3063	13	3492		6772
11	67		927	117	2373	10	1030		4524
12	328		2627	55	2786	9			5805
(blank)									
Grand Total	1991	456	5739	2192	32669	249	26681		69977



Sum of Amt Month	Books	Cafes	Entertainment	Fuel	Groceries	Music	Restaurants	(blank)	Grand Total
1	169	36	271	209	2147	15			2847
2	476	59	142	202	2820	15			3714
3	160	48	51	329	2348	46	2519		5501
4	418	34	307	100	2985	9	3299		7152
5	96	63	240	288	2911	14	2136		5748
6	38	145	309	198	2905	86	3352		7033
7	60	33	722	228	2834	6	3419		7302
8	79	38	143	138	3120	17	3651		7186
9	61			163	2377	9	3783		6393
10	39			165	3063	13	3492		6772
11	67		927	117	2373	10	1030		4524
12	328		2627	55	2786	9			5805
(blank)									
Grand Total	1991	456	5739	2192	32669	249	26681		69977



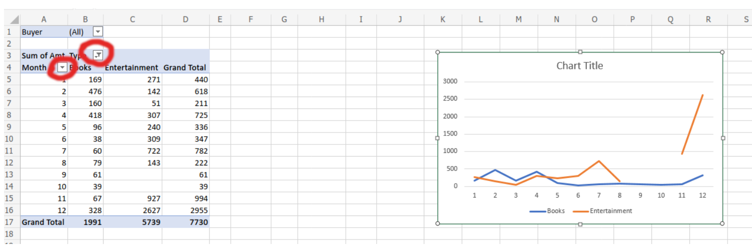
Sum of Amt Month	Books	Cafes	Entertainment	Fuel	Groceries	Music	Restaurants	(blank)	Grand Total
1	169	36	271	209	2147	15			2847
2	476	59	142	202	2820	15			3714
3	160	48	51	329	2348	46	2519		5501
4	418	34	307	100	2985	9	3299		7152
5	96	63	240	288	2911	14	2136		5748
6	38	145	309	198	2905	86	3352		7033
7	60	33	722	228	2834	6	3419		7302
8	79	38	143	138	3120	17	3651		7186
9	61			163	2377	9	3783		6393
10	39			165	3063	13	3492		6772
11	67		927	117	2373	10	1030		4524
12	328		2627	55	2786	9			5805
(blank)									
Grand Total	1991	456	5739	2192	32669	249	26681		69977

Exercise

Use file shopping.csv

Exercise: Can you plot (multiple line plot) the Books and Entertainment only?

- (hint: you can filter row labels and *column* labels.)



Sum of Amt Month	Books	Entertainment	Grand Total
1	169	271	440
2	476	142	618
3	160	51	211
4	418	307	725
5	96	240	336
6	38	309	347
7	60	722	782
8	79	143	222
9	61		61
10	39		39
11	67	927	994
12	328	2627	2955
Grand Total	1991	5739	7730

2 Processing Long Tables in MATLAB

groupsummary

https://au.mathworks.com/help/releases/R2020a/matlab/ref/ double.groupsummary.html

- groupsummary is one of the tools that MATLAB offers to obtain summaries from a long table.
- `groupsummary(T, groupvars, method, datavars)`
  - `T`: the table

- `groupvars`: the variables to group
- `method`: how to group them, e.g. `'sum'`, `'mean'`, etc. By default, if we don't say anything, it will count them.
- `datavars`: what column to apply the method to. All other columns are ignored. By default, if we don't say anything, ir will apply the method to all columns (except those specified in 'groupvars').

Demonstration
See file shopping.csv and script groupsummary_script.mlx

	1	2	3	4
	Date	Buyer	Type	Amt
1	01-Jan	Mom	Fuel	50
2	02-Jan	Mom	Groceries	120
3	03-Jan	Dad	Cafes	10
4	04-Jan	Dad	Fuel	40
5	04-Jan	Kelly	Groceries	129
6	05-Jan	Mom	Cafes	12

1. How would you find the *total* shopping of each buyer?

```
groupsummary(shopping,'Buyer','sum','Amt')
```

2. How would you find the *average* shopping of each buyer *per category*?

```
groupsummary(shopping,{'Buyer','Type'},'mean','Amt')
```

The live script `groupsummary_script.mlx` is an example of a well-documented script. When you write your scripts for the unit project, aim at writing with this level of detail.

groupsummary with binning
See file shopping.csv and script groupsummary_script.mlx
https://au.mathworks.com/help/releases/R2020a/matlab/ref/ double.groupsummary.html

- Sometimes we want to group by *parts of a date*. We can do this by specifying *group bins*.
- `groupsummary(T, groupvars, groupbins, method, datavars)`
- Possible types of group bins for dates:
  - `dayname`: the day of the week.
  - `monthname`: the month of the year.
  - `month`: by months.

1. How would you find the *total* shopping of *each month*?

```
groupsummary(shopping,'Date','monthname','sum','Amt')
```

Possible values for binning dates, times, and duration:

Value	Description	Data Type
'second'	Each bin is 1 second.	datetime and duration
'minute'	Each bin is 1 minute.	datetime and duration
'hour'	Each bin is 1 hour.	datetime and duration
'day'	Each bin is 1 calendar day. This value accounts for Daylight Saving Time shifts.	datetime and duration
'week'	Each bin is 1 calendar week.	datetime only
'month'	Each bin is 1 calendar month.	datetime only
'quarter'	Each bin is 1 calendar quarter.	datetime only
'year'	Each bin is 1 calendar year. This value accounts for leap days.	datetime and duration
'decade'	Each bin is 1 decade (10 calendar years).	datetime only
'century'	Each bin is 1 century (100 calendar years).	datetime only
'secondofminute'	Bins are seconds from 0 to 59.	datetime only
'minuteofhour'	Bins are minutes from 0 to 59.	datetime only
'hourofday'	Bins are hours from 0 to 23.	datetime only
'dayofweek'	Bins are days from 1 to 7. The first day of the week is Sunday.	datetime only
'dayname'	Bins are full day names such as 'Sunday'.	datetime only
'dayofmonth'	Bins are days from 1 to 31.	datetime only
'dayofyear'	Bins are days from 1 to 366.	datetime only
'weekofmonth'	Bins are weeks from 1 to 6.	datetime only
'weekofyear'	Bins are weeks from 1 to 54.	datetime only
'monthname'	Bins are full month names such as 'January'.	datetime only
'monthofyear'	Bins are months from 1 to 12.	datetime only
'quarterofyear'	Bins are quarters from 1 to 4.	datetime only

unstack

Use file weather_data.csv

https://au.mathworks.com/help/releases/R2020a/matlab/ref/unstack.html

- unstack can be used to convert a long table into a wide table.
- `unstack(S,vars,ivar)`
  - `S` The table to unstack.
  - `vars` The values to fill in the new columns.
  - `ivar` The indicator variables. Different values in these variables will generate separate columns in the resulting table.

Example

```
S = readtable('weather_data.csv');
S2 = removevars(S,"Var1");
T = unstack(S2,'data','param');
```

218454×5 table

	date	siteId	MaximumTemper...	MinimumTemper...	Precipitation
1	2003-01-01	'ACRE'	-3.7900	-7.9400	0
2	2003-01-02	'AlbertLea'	-3.6500	-5.9600	0
3	2003-01-03	'Ames'	-1.0900	-12.3900	11.3200
4	2003-01-04	'Antigo'	-1.0400	-12.1500	0
5	2003-01-05	'Appleton'	-1.1200	-4.0300	3.0400
6	2003-01-06	'Arlington'	-0.1200	-3.2400	0.4900
7	2003-01-07	'BeanBeat'	-1.0300	-5.2400	0
8	2003-01-08	'Brookings'	9.1400	-2.5000	0
9	2003-01-09	'Brownstown'	10.7400	-1.3100	0
10	2003-01-10	'Columbia'	1.3800	-10.1800	0
11	2003-01-11	'Crookston'	-9.2200	-17.1700	0
12	2003-01-12	'Dekalb'	-9.3900	-17.8800	0

According to the MATLAB documentation:

> The unstack function treats the remaining variables differently in tables and timetables.
>
> - If S is a table, then unstack treats the remaining variables as grouping variables. Each unique combination of values in the grouping variables identifies a group of rows in S that is unstacked into a single row of U.
> - If S is a timetable, then unstack discards the remaining variables. However, unstack treats the vector of row times as a grouping variable.

In our example, we are unstacking a table (not a timetable), and as a result we need to remove the first column with name "Var1". If we don't remove this column, since all values of "Var1" are unique, unstack will not merge the rows correctly.

Try it out. Do not remove "Var1" and observe that most values of the resulting table are "NaN". "NaN" means "not a number", and it basically means (in this context) that the cells with "NaN" values are empty.

3 Data Analysis in MATLAB

Analysing the Data

- Excel and MATLAB provide various tools for data analysis.
- Some of these tools are in Excel's *Data Analysis Tool Pack* plug-in.
  - *https://support.office.com/en-us/article/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4*
  - Available in the desktop version, sometimes in the online version (not available at Macquarie University).
  - Understanding most of these tools is beyond the scope of this unit.
- In this lecture we will look at how to do the following in MATLAB:
  - Plotting data.
  - Finding correlations.

3.1 Plotting in MATLAB

Plotting with MATLAB

Demo using file trees.csv

MATLAB offers several options to display scatterplots (and other plots):

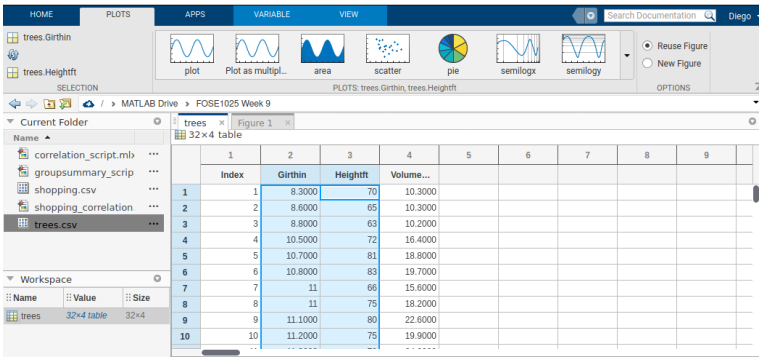
1. Executing the command (in the command window or in a script), e.g. for a scatterplot:

```
scatter(trees.Girthin,trees.Heightft)
```

2. Interacting graphically (more intuitive; see demo in the lecture)
   - This allows you to do more complex plots, e.g. multiple charts.
   - After interacting with MATLAB, you will see the resulting MATLAB command in the command window.

Steps to plot using the interactive interface

1. Double click on the variable that contains the table.
2. Select the columns to plot.
3. At the "Plots" tab, select the desired plot type.



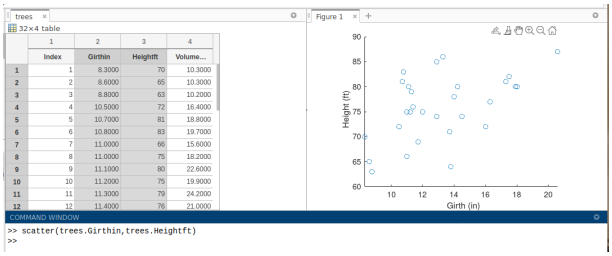
3.2 MATLAB: Finding Correlations

What is Correlation?

- Sometimes two separate sources of information are measuring the same property.
- You may detect this by observing that the values are the same.
- But sometimes this is not the case:
  - Each source may use different units of measure (e.g. metric vs. imperial).
  - Each source makes an independent measure that has some noise.
- In other cases, two variables are correlated but might not be identical.
  - For example, tree trunk height and girth are correlated.
  - Taller trees will normally have thicker trunks.
- MATLAB (and Excel) can detect the degree of *correlation* between two series of numbers.

Finding Correlations Graphically
Screenshot using file trees.csv

- A *scatterplot* can plot one variable against the other.
- If the two variables are not correlated, the scatterplots will look random.
- If the scatterplot has a distinct shape, the two variables are correlated.
- For example, if the shape looks like a line, then the two variables have a *linear correlation.*



In this scatterplot we can see that the tree height and girth are loosely correlated, since the plot has a rising trend.

Finding Correlations on Multiple Columns

Examples and explanations in correlation_script.mlx using file trees.csv

- Scatterplots are intuitive but may be cumbersome if you want to check the correlations among many columns.
  - E.g. if there are 10 columns you will need to make a plot for each possible pair.
  - This means making $10 \times 9 = 90$ plots.
- MATLAB's `corr` computes *Pearson's Linear Correlation Coefficient* but you can specify others.
  - e.g. `corr(trees.Girthin,trees.Heightft)` computes the correlation between columns `Girthin` and `Heightft` of table `trees`.
  - `corr(trees.Girthin,trees.Heightft,"rows","complete")` will ignore empty values.
- A number close to 1 (or -1) indicates positive (or negative) correlation; 0 means no correlation.

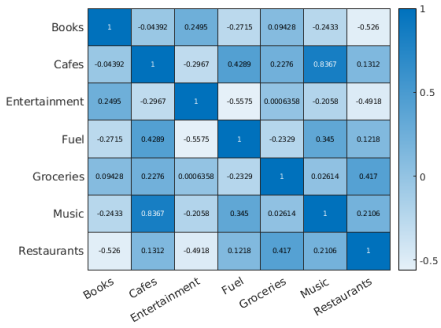
Correlation Matrix

Examples and explanations in correlation_script.mlx

- MATLAB's `corrcoef` function returns a *correlation matrix*.
  - A correlation matrix returns the pairwise correlation between multiple columns.
- The input to `corrcoef` must be a matrix (not a table).
  - You can use the notation `mytable{rows,columns}` to extract rows and columns from a table and generate a matrix that can be fed to `corrcoef`
  - *https://www.mathworks.com/help/matlab/matlab_prog/access-data-in-a-table.html*
- This matrix can then be displayed using a *heatmap*.

Exercise
See detailed solution in script shopping_correlation.mlx

- File: shopping.csv
- Build the correlation matrix between all types of shopping.
- What are the two most correlated types of shopping?
- Show it clearly by creating a heatmap.



The script is an example of a well-documented script. When you submit your scripts, try to produce the same level of detail.

Take-home Messages

- EXCEL: You must be able to use pivot tables for a range of tasks.
- EXCEL: You must be able to create charts based on pivot tables.
- EXCEL and MATLAB: You must be able to plot data.
- MATLAB: You must be able to detect whether two variables are correlated.
- MATLAB: You should also be able to display correlation using heatmaps.

What's Next

- *Test 3 in this week's SGTA.*
- *No classes Monday 4 October — Happy holiday!*
- Week 10 lecture: Ethics related to Scientific Computing.
- Week 11:
  - Wed 20 October: Submit the project.
  - Fri 22 October: Submit Collaborator employability hurdle.