FOSE1025 — Scientific Computing

Week 9 Lecture 1: Summarising and Analysing Data

Diego Mollá

FOSE1025 2020H2

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.

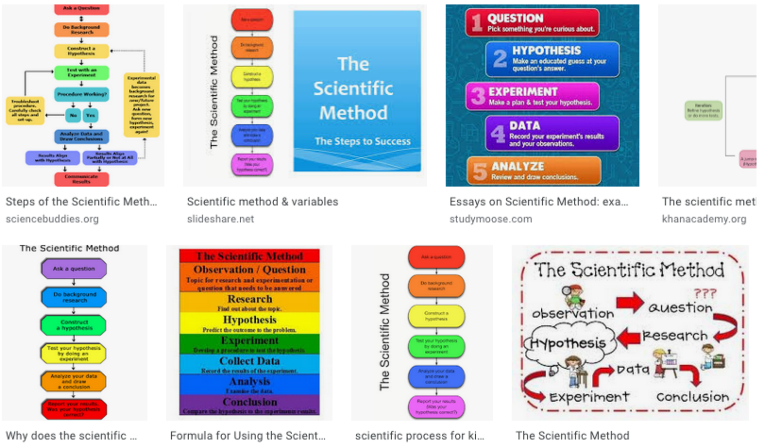
Update October 2, 2020

Contents

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


Steps of the Scientific Meth...
sciencebuddies.org
Scientific method & variables
slideshare.net
Essays on Scientific Method: exa...
studymoose.com
The scientific met
khanacademy.org
Why does the scientific ...
Formula for Using the Scient...
scientific process for ki...
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

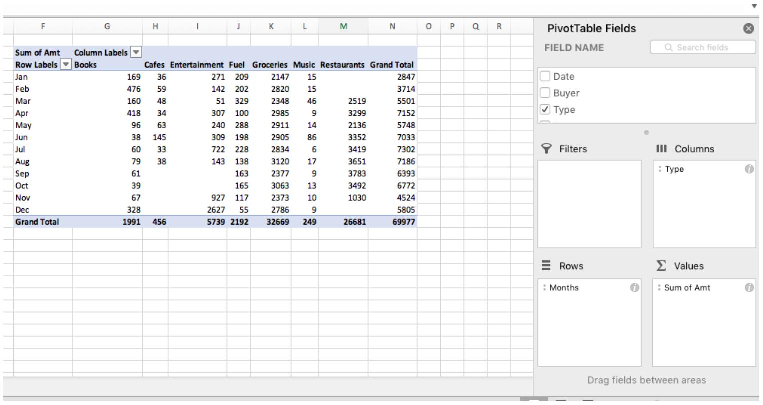
We are covering these aspects in FOSE1025:

- Import data from external files (e.g. CSV) — Week 3.
- Explore the data — Week 4, 5.
- Clean the data — Week 7.
- Preprocess, transform the data — Week 8.
  - aka "data wrangling," "data munging".
- *Analyse, summarise, interpret the data — Week 5, Week 9.*

1 Excel's Pivot Tables for Charts

A Simple Pivot Table

Use file shopping.csv



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

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.

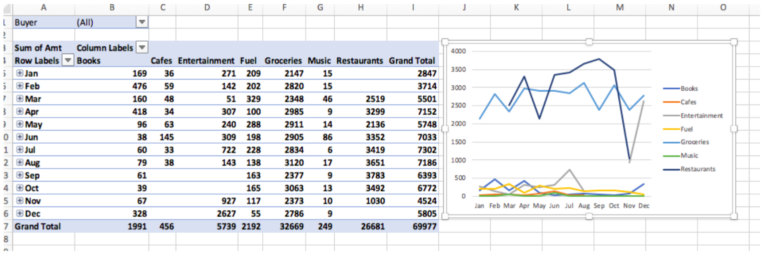
Values

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

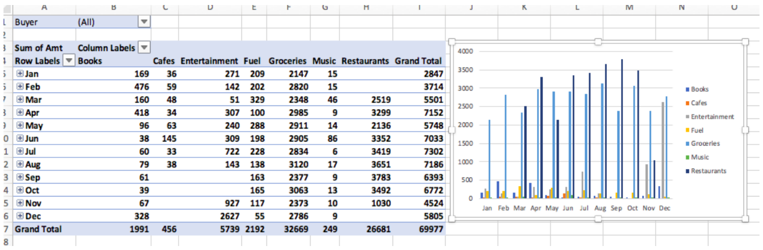
Pivot Tables for Charts

Use file shopping.csv

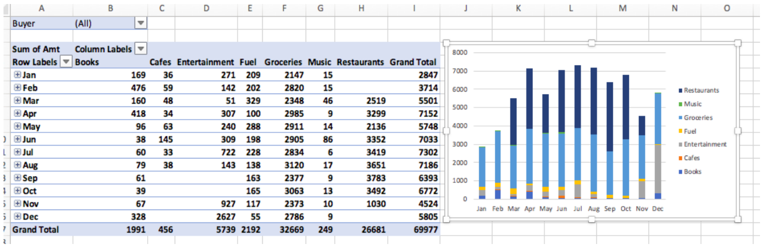
- Pivot tables facilitate the transformation of data for the creation of complex plots.
- In a *multiple chart*, each column of a table is plotted overlayed with the rest. Good for line plots.
- In a *clustered chart*, each row forms a cluster. Good for bar charts.
- In a *stacked chart*, columns of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



Buyer	(All)							
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



Buyer	(All)							
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



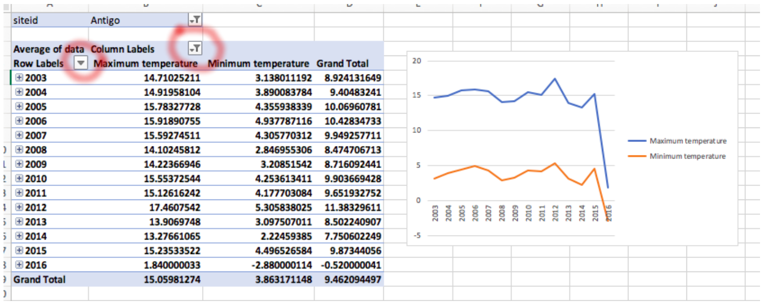
Buyer (All)

Sum of Amt Row Labels	Column 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

Pivot Charts: Pivot Tables *and* Charts!

Use file weather_data.csv

- Pivot tables are so useful for making charts that there's a tool for that combines both: Pivot charts!
- Exercise: Can you plot (multiple line plot) the maximum and minimum temperature of Antigo as it changes over time? Do not plot precipitation.
  - (hint: you can filter row labels and *column* labels.)



siteid Antigo

Average of data Row Labels	Column Labels Maximum temperature	Minimum temperature	Grand Total
⊞ 2003	14.71025211	3.138011192	8.924131649
⊞ 2004	14.91958104	3.890083784	9.40483241
⊞ 2005	15.78327728	4.355938339	10.06960781
⊞ 2006	15.91890755	4.937787116	10.42834733
⊞ 2007	15.59274511	4.305770312	9.949257711
⊞ 2008	14.10245812	2.846955306	8.474706713
⊞ 2009	14.22366946	3.20851542	8.716092441
⊞ 2010	15.55372544	4.253613411	9.903669428
⊞ 2011	15.12616242	4.177703084	9.651932752
⊞ 2012	17.4607542	5.305838025	11.38329611
⊞ 2013	13.9069748	3.097507011	8.502240907
⊞ 2014	13.27661065	2.22459385	7.750602249
⊞ 2015	15.23533522	4.496526584	9.87344056
⊞ 2016	1.840000033	-2.880000114	-0.520000041
Grand Total	15.05981274	3.863171148	9.462094497

2 Processing Long Tables in MATLAB

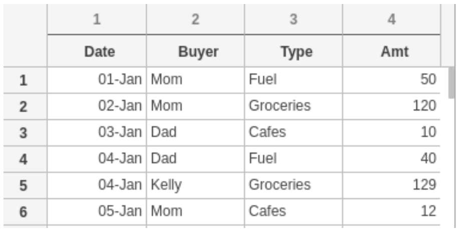
groupsummary

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 total shopping of each buyer per category?

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

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

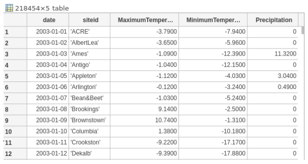
Use file weather_data.csv

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	'Bean&Beet'	-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

Participation Task

- Complete the exercises of MATLAB Grader (Lecture Participation Week 9).
- Participation will count if your submission passes at least 1 test in each exercise.

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*
  - 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 and showing trends.
  - Finding correlations.

3.1 MATLAB: Plotting with Trends

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:

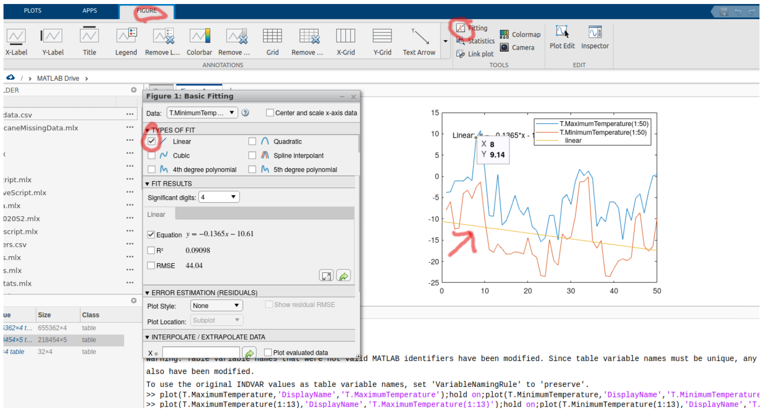
   ```
   scatter(trees.Girthin, trees.Heightft)
   ```

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.

Adding a Trend Line

Screenshot using file weather_data.csv

- Once a plot is made, we can fit lines using the ”fitting” dialogue.
- For example, try to add a trend line by selecting a ”linear” type of fit (see screenshot).



ue	Size	Class
5362×4 t...	655362×4	table
8454×5 t...	218454×5	table
:4 table	32×4	table

Figure 1: Basic Fitting

Data: T.MinimumTemp... ▾ □ Center and scale x-axis data

▾ TYPES OF FIT

☑ Linear □ Quadratic
□ Cubic □ Spline interpolant
□ 4th degree polynomial □ 5th degree polynomial

▾ FIT RESULTS

Significant digits: 4 ▾

Linear

☑ Equation $y = -0.1365x - 10.61$

□ R² 0.09098

□ RMSE 44.04

▾ ERROR ESTIMATION (RESIDUALS)

Plot Style: None ▾ □ Show residual RMSE

Plot Location: Subplot ▾

▾ INTERPOLATE / EXTRAPOLATE DATA

X = □ Plot evaluated data

```
Warning: Table variable names that were not valid MATLAB identifiers have been modified. Since table variable names must be unique, any
also have been modified.
To use the original INDVAR values as table variable names, set 'VariableNamingRule' to 'preserve'.
>> plot(T.MaximumTemperature,'DisplayName','T.MaximumTemperature');hold on;plot(T.MinimumTemperature,'DisplayName','T.MinimumTemperature
>> plot(T.MaximumTemperature(1:13),'DisplayName','T.MaximumTemperature(1:13)');hold on;plot(T.MinimumTemperature(1:13),'DisplayName','T.
```

3.2 MATLAB: Finding Correlations

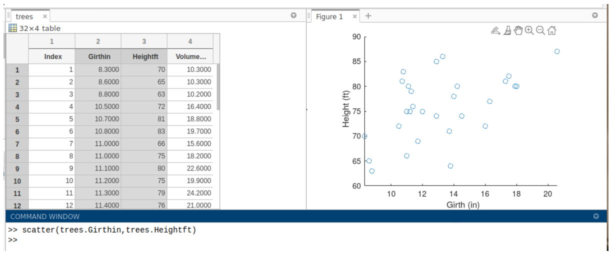
What is Correlation?

- 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 (e.g. as in this year's project).
- 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.*



trees

32×4 table

	1	2	3	4
	Index	Girthin	Heightft	Volume...
1	1	8.3000	70	10.3000
2	2	8.6000	65	10.3000
3	3	8.8000	63	10.2000
4	4	10.5000	72	16.4000
5	5	10.7000	81	18.8000
6	6	10.8000	83	19.7000
7	7	11.0000	66	15.6000
8	8	11.0000	75	18.2000
9	9	11.1000	80	22.6000
10	10	11.2000	75	19.9000
11	11	11.3000	79	24.2000
12	12	11.4000	76	21.0000

COMMAND WINDOW

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

Finding Correlations on Multiple Columns

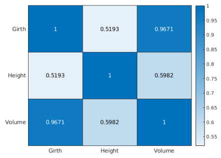
Examples and explanations in correlation_script.mlx

- 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 10×9 = 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

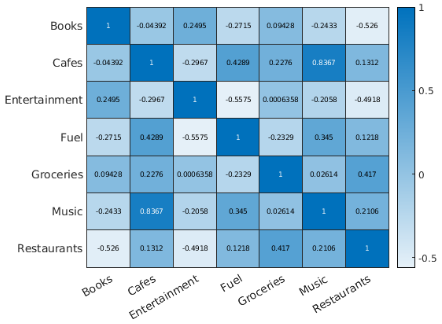
- MATLAB's **corrcoef** function returns a *correlation matrix.*
  - A correlation matrix returns the pairwise correlation between multiple columns.
- The input to **corrcoef** must be a matrix (not a table).
  - You can use the notation mytable{rows,columns} to extract rows and columns from a table 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.



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 or MATLAB: You must be able to show trends by adding trend lines to a plot.
- MATLAB: You must be able to detect whether two variables are correlated.
- MATLAB: Ideally, you should also be able to display correlation using heatmaps.

What's Next

- Week 10 lecture: Ethics related to Scientific Computing.
- Week 11:
  - Wed 21 October: Submit the project.
  - Fri 23 October: Submit Collaborator employability hurdle.