FOSE1025 — Scientific Computing

Week 6 Lecture 1: Towards Using Scripts for Reproducibility

Diego Mollá

Department of Computer Science
Macquarie University

FOSE1025 2020H2

Announcements

1. Liaison meeting on Week 7.
2. Quiz 2 during this week's practical 1 (Friday evening for external students).

   Internal: Submit as an assignment.
   External: Submit as a quiz.
3. Lecture notes and activity slides in Echo360.
4. Remember to register to MATLAB Grader course.
5. Material in github (next slide).

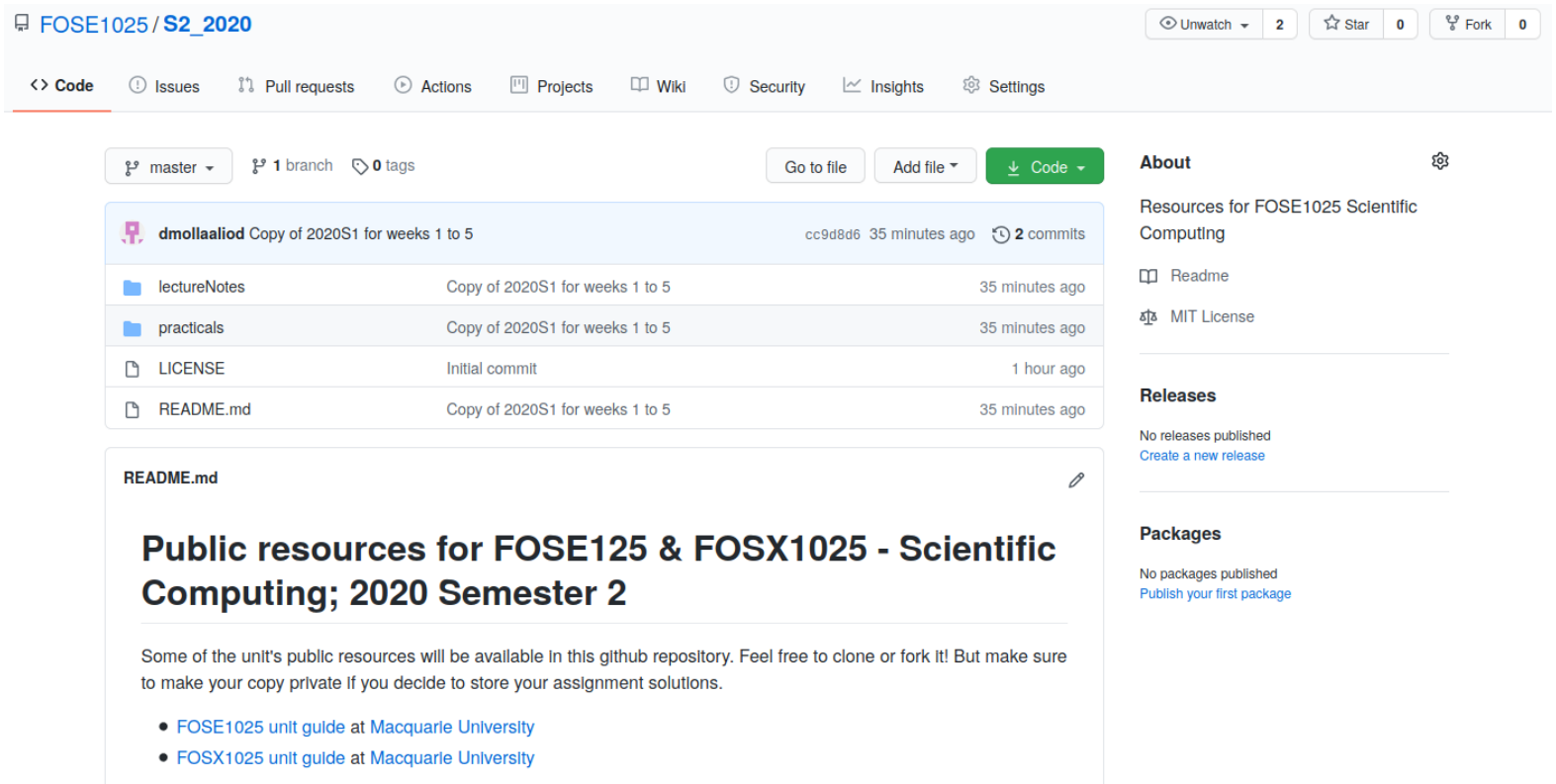
Programme

1. Review: Excel for Science
2. Scripts for Reproducibility
3. MATLAB

Reading

- Lecture notes
- https://au.mathworks.com/help/matlab/getting-started-with-matlab.html
- https://au.mathworks.com/videos/getting-started-with-matlab-1564521672719.html

FOSE1025's public github page



Programme

The Scientific Method

Steps of the Scientific Meth...
sciencebuddies.org

Scientific method & variables
slideshare.net

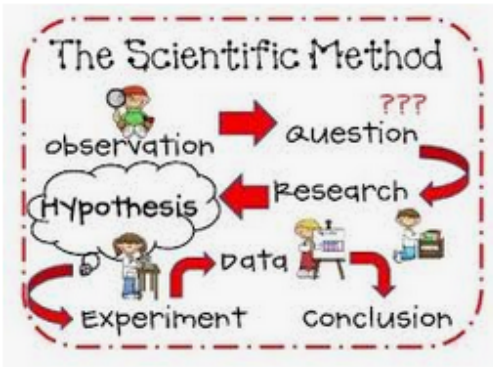
Essays on Scientific Method: exa...
studymoose.com

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, Week 5.
- (you are here)
- Clean the data — Week 7.
- Preprocess, transform the data — Week 8.
- Analyse, summarise, interpret the data — Week 5, Week 9.

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
- ...

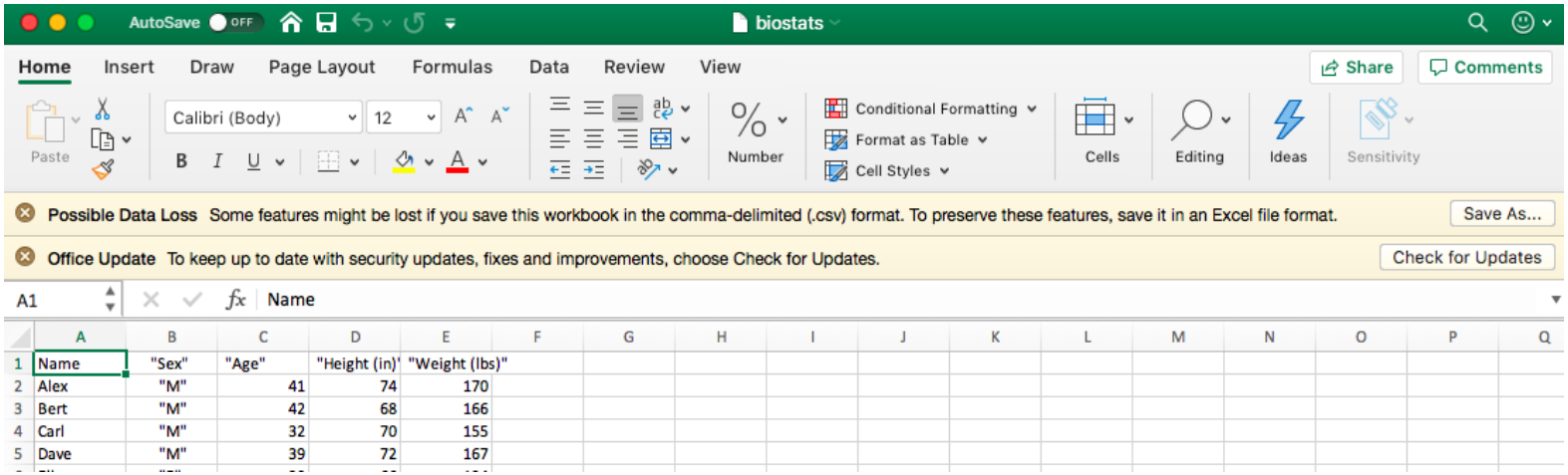
Example CSV File

(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
```

Careful if you double-click on a CSV file!



- If you double-click on a CSV file, Excel will open the file.
- But the file opened is a CSV file, not an Excel (.xlsx) file!
  - Read the warning that you get if you double-click on the CSV file.
- There are many things that you cannot save in a CSV file.
  - Formulas, formatting, charts, etc.

Tables in Excel

- 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	F
1	Name	"Sex"	"Age"	"Height (i	"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	

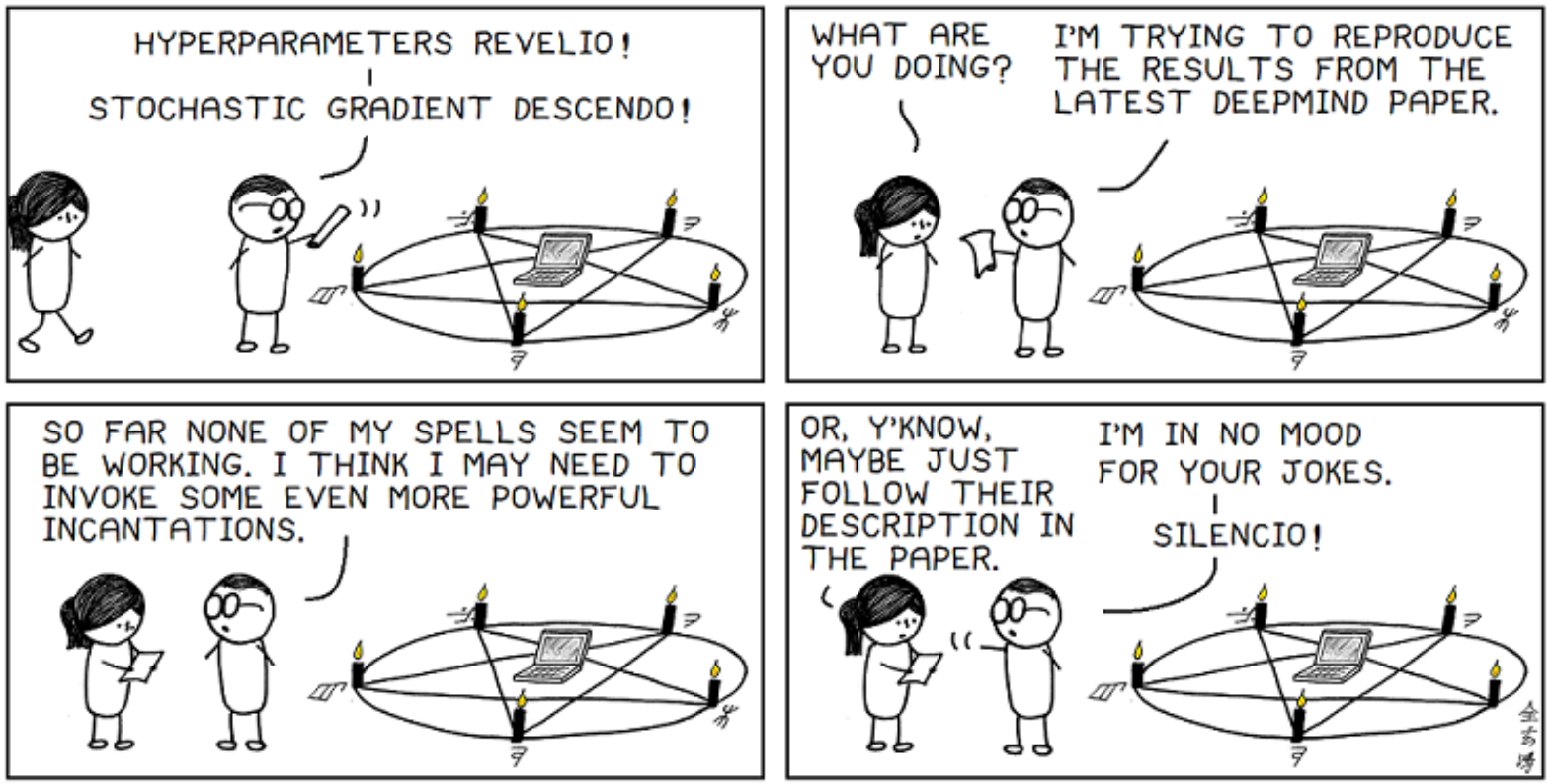
Question: What are the data types of each column?

Programme

The Problem with Reproducibility

It can be difficult to write clearly enough to allow reproducibility.

https://abstrusegoose.com/588



Reproducibility in Science

- When you conduct science, you need to make sure that others can reproduce what you did.
  - If others can reproduce what you did, then your claims are more likely to be taken as valid.
- Reproducibility means that someone else should be able to do the same as you did by following your instructions.
- When the experiments are performed with computers, reproducibility can mean one of two:
  1. "I can re-implement what you did after I read your report."
  2. "I can run the code that you wrote."
- The employability modules ("Achiever" and "Communicator") touch item 1.
- Here we will touch item 2.

Scripting Languages

- Scripting languages are programming languages designed for rapid prototyping.
  - $\Rightarrow$ These languages make it easy to quickly write and execute a program.
- Scripting languages are normally interpreted languages.
  - $\Rightarrow$ This means that you can execute instructions one by one using a run time environment.

Example of Steps

1. Start the run time environment (e.g. MATLAB).
2. Type instructions (or load instructions stored in a file).
3. Run the instructions in the run time environment.

Top 10 Programming Languages for Data Science

https://www.analyticsinsight.net/top-10-data-science-programming-languages-for-2020/

1. Python (popular among programmers and web developers)
2. R (popular among statisticians)
3. SQL (designed for querying relational databases)
4. C (C++)
5. Java
6. JavaScript (originally designed to run in a browser)
7. MATLAB (the focus of this unit)
8. Scala
9. Swift
10. Julia

Demonstration Using MATLAB Online

- In this demonstration, the runtime runs in the cloud.
- We use a web browser to interact with the runtime.
- Can be done with any computer as long as it has:
  - An internet connection.
  - A modern browser.
- There is no need to install additional software in your computer.

MATLAB Online

- https://au.mathworks.com/academia/tah-portal/macquarie-university-916052.html
- Create an account with your student email address
- Do not use your student password (create a new one)

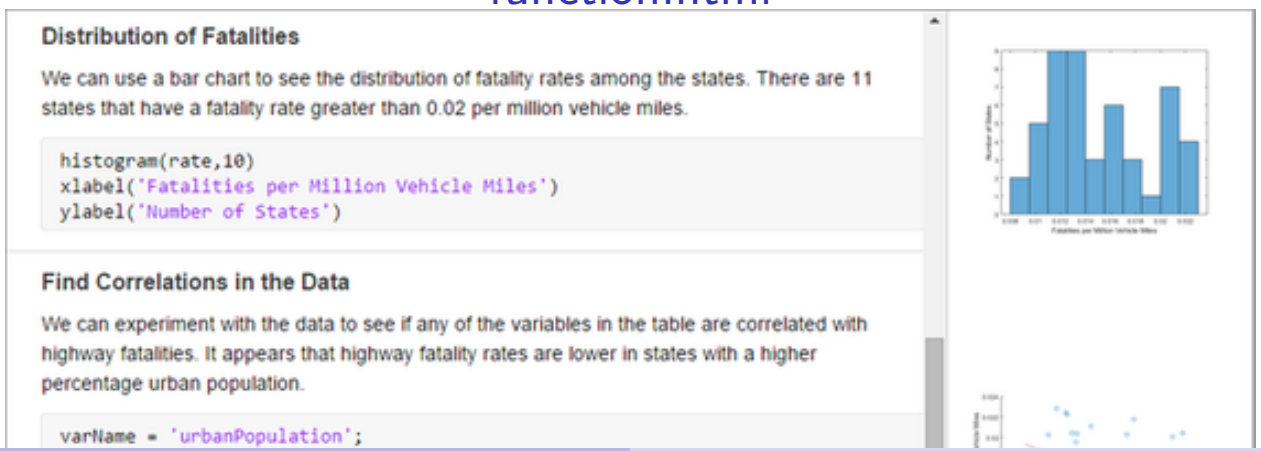
Scripting Languages and Reproducibility

- Instructions written in a scripting language ensure reproducibility . . . or does it not?
- While instructions written in a scripting language can be executed by a computer . . .
  - . . . instructions may not do what we intended them to do (e.g. because there are errors in the instructions).
  - Poorly-written scripts may not be understandable by people
    $\Rightarrow$ and then we cannot tell if they are correct.
  - Portability: Scripts running in a computer might not run in another computer.
    $\Rightarrow$ often you need to provide instructions for installation of necessary software dependencies.
- Normally we want to supplement the instructions with comments and explanations.

Notebooks for Reproducibility

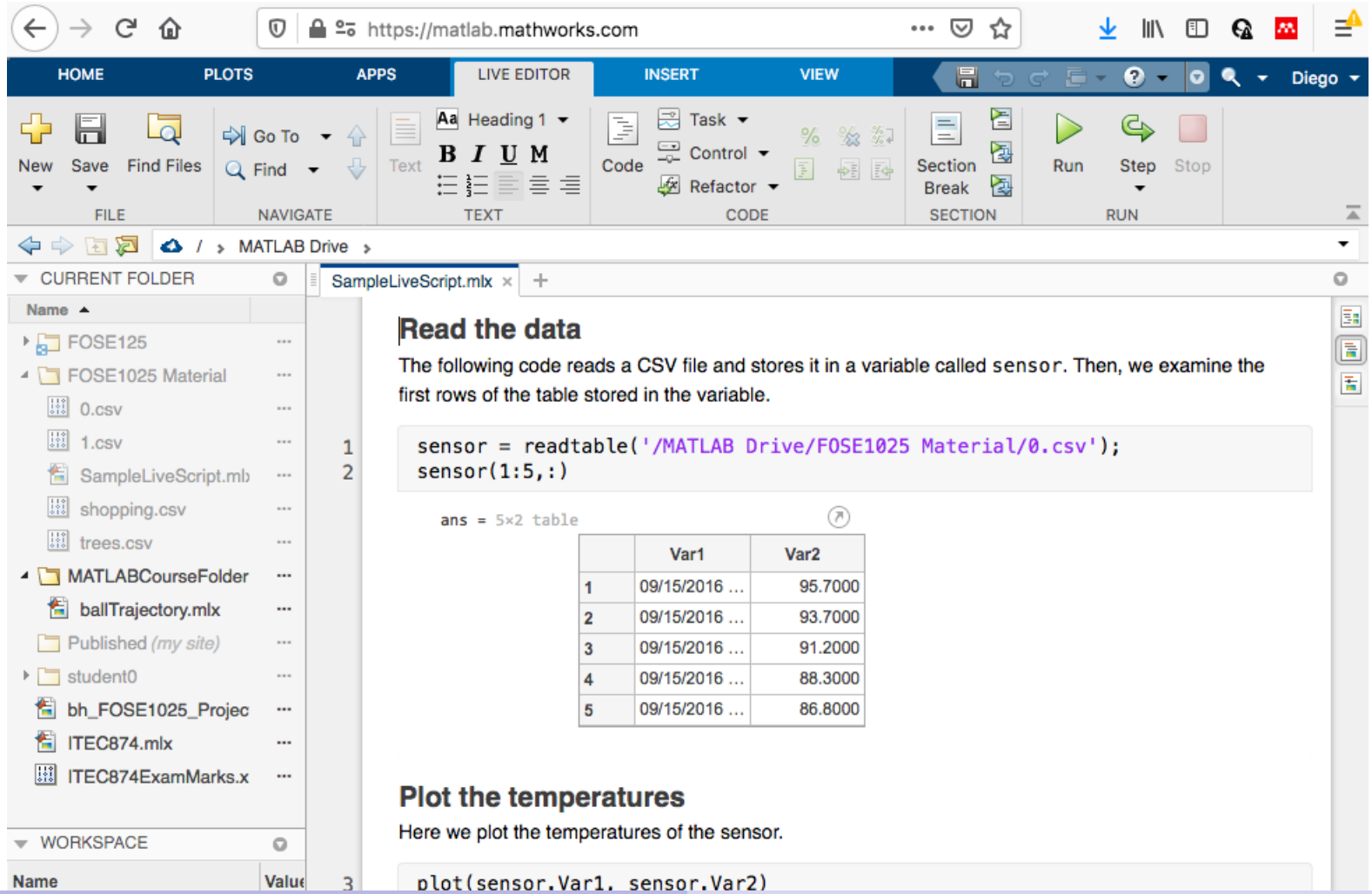
- Some run time environments allow the creation of notebooks.
  - Called live scripts in MATLAB.
- These notebooks are the digital equivalent of lab notebooks.
- Notebooks contain sections that can be executed.
- The results of execution appear in the notebook.
- Notebooks also contain formatted text for documentation and explanations.

https://au.mathworks.com/help/matlab/matlab_prog/what-is-a-live-script-or-function.html



Demonstration of a Live Script

SampleLiveScript.mlx



Programme

What is MATLAB?

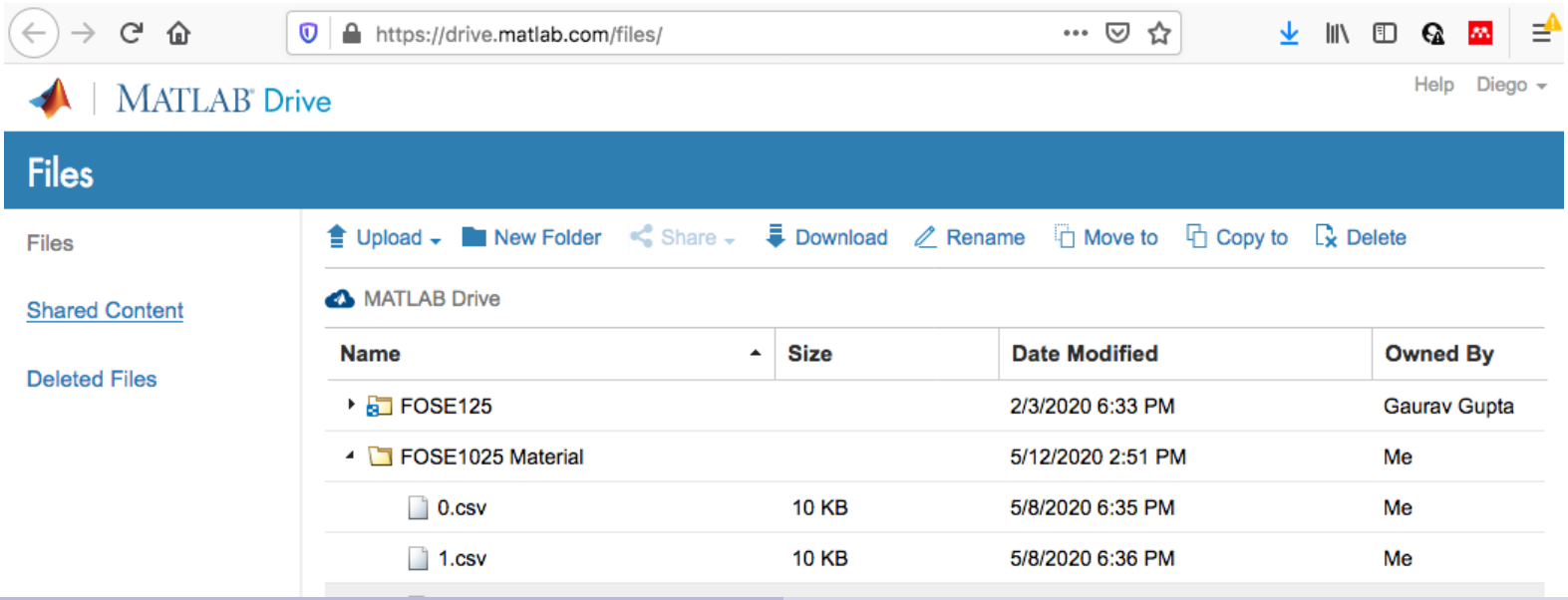
- MATLAB is a scripting language.
- Includes types designed to store and manipulate data.
  - Matrices (MATLAB = MATrix LABoratory)
  - Tables (our focus in this unit)
- Includes a large library of functions for data analysis, manipulation, and visualisation.
- Has extensive documentation and on-line courses.
- Easy to use
- Others programming languages have attempted to integrate some of MATLAB's features.
  - Matrices, tables
  - Plots
  - Interactive notebooks

Accessing MATLAB and MATLAB Online

- Macquarie University has a license for students: https://au.mathworks.com/academia/tah-portal/macquarie-university-916052.html
- MATLAB Online here: https://matlab.mathworks.com/
- Getting started: https://au.mathworks.com/help/matlab/getting-started-with-matlab.html
- Self-paced courses: https://matlabacademy.mathworks.com/

MATLAB Online and MATLAB Drive

- MATLAB Online runs in the cloud.
- To upload files to the cloud you can use MATLAB Drive.
- You can use a browser to upload and download files.
- Or you can install software that integrates with your computer file system.
  - It looks and feels like MATLAB drive is a folder in your computer.



Loading data in MATLAB

- MATLAB Fundamentals, Chapter 10, "Tables of Data"
- https://au.mathworks.com/help/releases/R2019b/matlab/matlab_prog/create-a-table.html
- MATLAB can store tables into variables.
- You can use the MATLAB "Import Data" wizard.
  - Looks like a more sophisticated version of Excel's Import tools.
- Or you can use the readtable instruction.
  - `trees = readtable("trees.csv");`

Processing data in MATLAB

- https://au.mathworks.com/help/releases/R2019b/matlab/matlab_prog/access-data-in-a-table.html
- Accessing a column: `girth = trees.("Girth (in)")`
- Accessing a full row: `sample = trees(5,:)`
- Adding a column:
  `trees.("Girth (cm)") = trees.("Girth (in)") * 2.54`
- Concatenating tables:

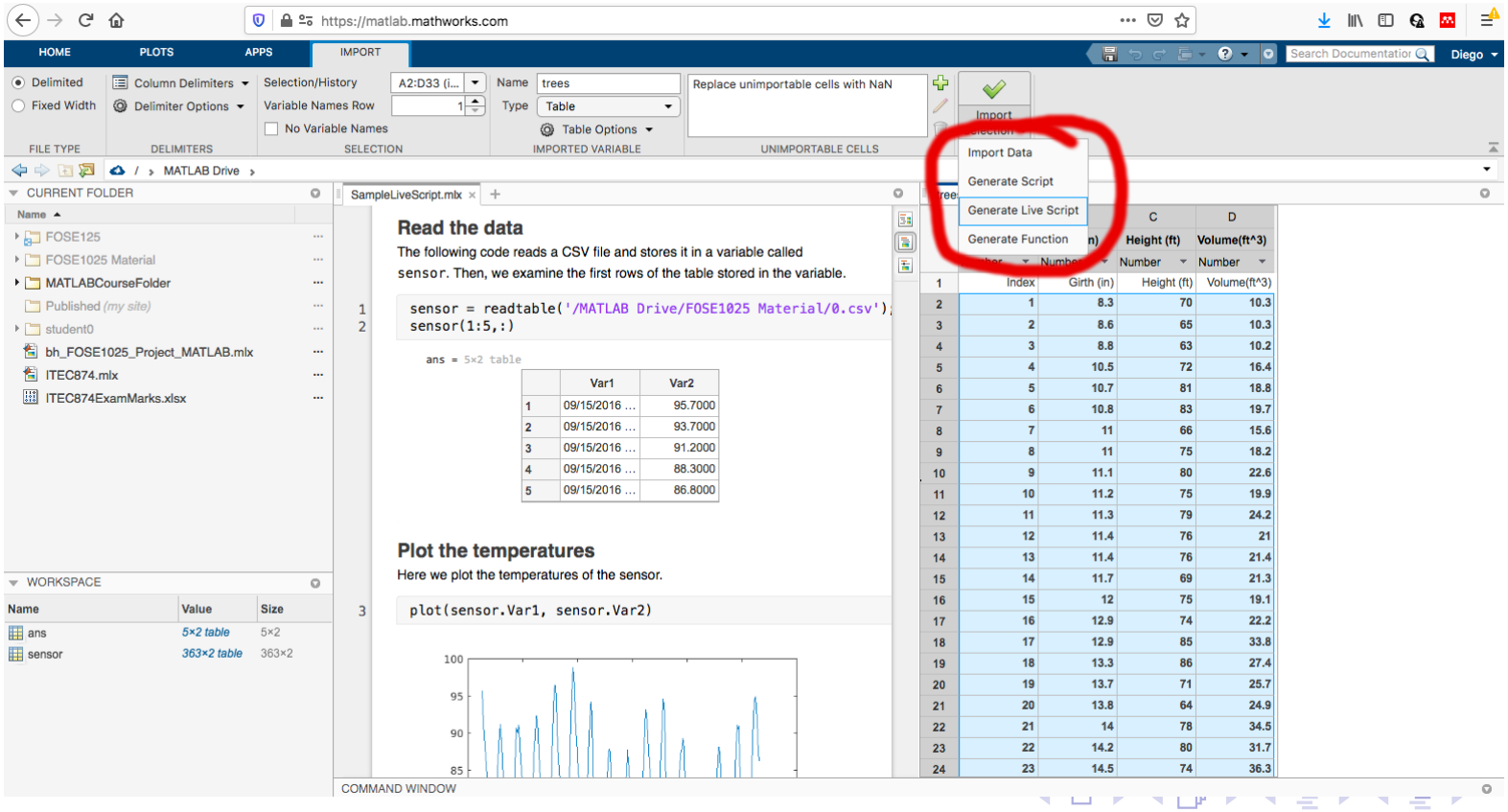
```
table0 = readtable("0.csv");
table1 = readtable("1.csv");
table = [table0; table1];
```

Participation Activity – MATLAB Grader

1. Access course FOSE1025 and FOSX1025 2020 S2 at MATLAB Grader
   - You should have received an invitation by email — check your student email.
2. Complete Lecture Participation Week 6
   - Read a CSV file
   - Extract a column
   - Extract a row

Creating and Reusing MATLAB Scripts

- Many MATLAB wizards can generate scripts.
- You can write your own script.
- Then you can run it again later.



Saving a MATLAB table as a CSV file

- MATLAB's `writetable` will write a table into a CSV file.
- https://au.mathworks.com/help/releases/R2019b/matlab/matlab_prog/create-a-table.html
- https://au.mathworks.com/help/matlab/ref/writetable.html
- `writetable(table, 'table.csv')` will save the table into file `table.csv`
- `writetable(table, 'table.csv', 'WriteRowNames', true)` will also write the column names.

Take-home Messages

- Excel as a tool to manage data in science.
- Excel tables.
- Scripting languages are powerful means to allow reproducibility.
- Scripting languages can be executed by a computer.
- Some environments allow the use of interactive notebooks for better reproducibility.
- MATLAB is a powerful scripting language designed for data analysis.
- Importing data in MATLAB.
- Accessing table rows and columns in MATLAB.

What's Next

- Week 6, practical 1: Quiz 2
- Week 7 lecture: Cleaning data
- Week 7, Friday 11 September: Communicator hurdle