FOSE1025 — Scientific Computing

Week 6 Lecture 1: Towards Using Scripts for Reproducibility

Diego Mollá

FOSE1025 2020H2

Abstract

In this lecture we will have a very brief introduction to the use of scripts to store and manipulate data. We move away from Excel and enter the area of programming. The emphasis will be on how to use scripts for reproducibility, and we will focus in a particular environment: MATLAB.

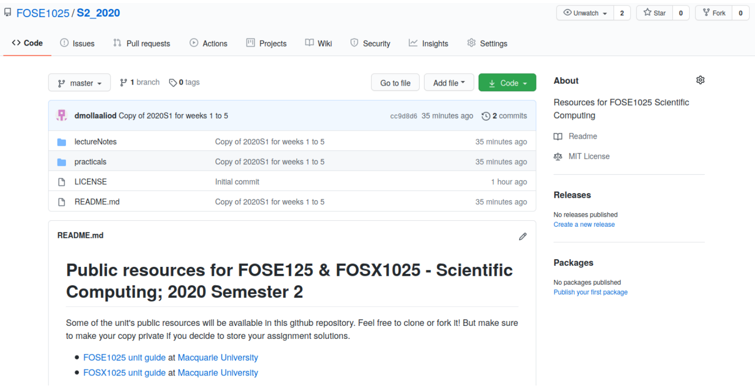
Update September 2, 2020

Contents

Reading

- These 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



dmollaaliod Copy of 2020S1 for weeks 1 to 5 cc9d8d6 35 minutes ago 2 commits

lectureNotes	Copy of 2020S1 for weeks 1 to 5	35 minutes ago
practicals	Copy of 2020S1 for weeks 1 to 5	35 minutes ago
LICENSE	Initial commit	1 hour ago
README.md	Copy of 2020S1 for weeks 1 to 5	35 minutes ago

Resources for FOSE1025 Scientific Computing

No releases published
Create a new release

No packages published
Publish your first package

README.md

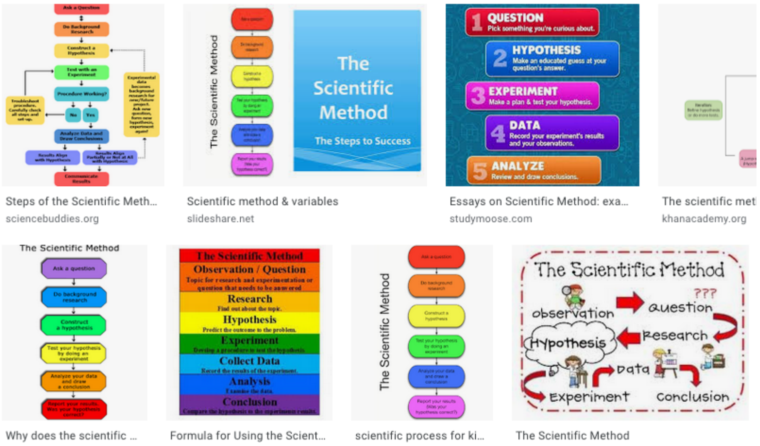
Public resources for FOSE125 & FOSX1025 - Scientific Computing; 2020 Semester 2

Some of the unit's public resources will be available in this github repository. Feel free to clone or fork it! But make sure to make your copy private if you decide to store your assignment solutions.

- FOSE1025 unit guide at Macquarie University
- FOSX1025 unit guide at Macquarie University

1 Review: Excel for Science

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
- . . .

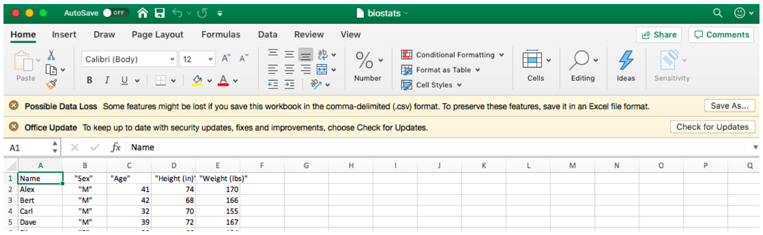
Example CSV File

(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
”Omar”,       ”M”,   38,       70,      145
”Page”,       ”F”,   31,       67,      135
”Quin”,       ”M”,   29,       71,      176
”Ruth”,       ”F”,   28,       65,      131
```

Careful if you double-click on a CSV file!



Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format. Save As...

Office Update To keep up to date with security updates, fixes and improvements, choose Check for Updates. Check for Updates

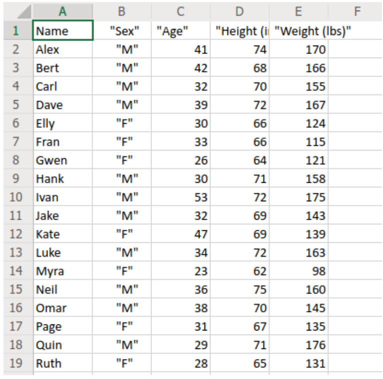
A1 *fx* Name

	A	B	C	D	E
1	Name	"Sex"	"Age"	"Height (in)'	"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

- 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	

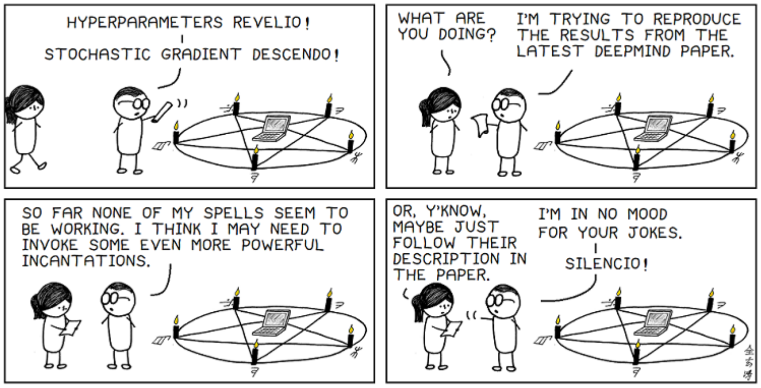
Question: What are the data types of each column?

2 Scripts for Reproducibility

The Problem with Reproducibility

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

https://abstrusegoose.com/588


HYPERPARAMETERS REVELIO!
STOCHASTIC GRADIENT DESCENDO!
WHAT ARE YOU DOING?
I'M TRYING TO REPRODUCE THE RESULTS FROM THE LATEST DEEPMIND PAPER.
SO FAR NONE OF MY SPELLS SEEM TO BE WORKING. I THINK I MAY NEED TO INVOKE SOME EVEN MORE POWERFUL INCANTATIONS.
OR, Y'KNOW, MAYBE JUST FOLLOW THEIR DESCRIPTION IN THE PAPER.
I'M IN NO MOOD FOR YOUR JOKES.
SILENCIO!

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*.

  ⇒ These languages make it easy to quickly write and execute a program.

- Scripting languages are normally *interpreted languages*.

  ⇒ 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

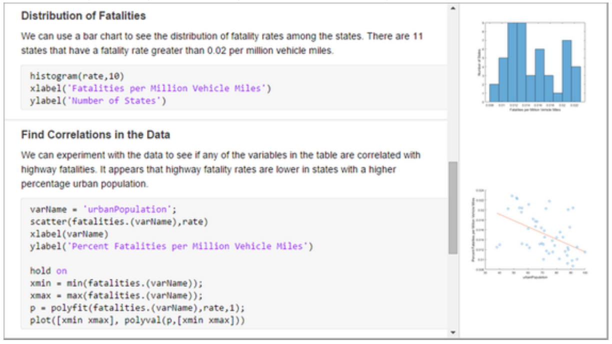
    ⇒ and then we cannot tell if they are correct.
  - *Portability:* Scripts running in a computer might not run in another computer.

    ⇒ 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



Distribution of Fatalities

We can use a bar chart to see the distribution of fatality rates among the states. There are 11 states that have a fatality rate greater than 0.02 per million vehicle miles.

```
histogram(rate,10)
xlabel('Fatalities per Million Vehicle Miles')
ylabel('Number of States')
```

Find Correlations in the Data

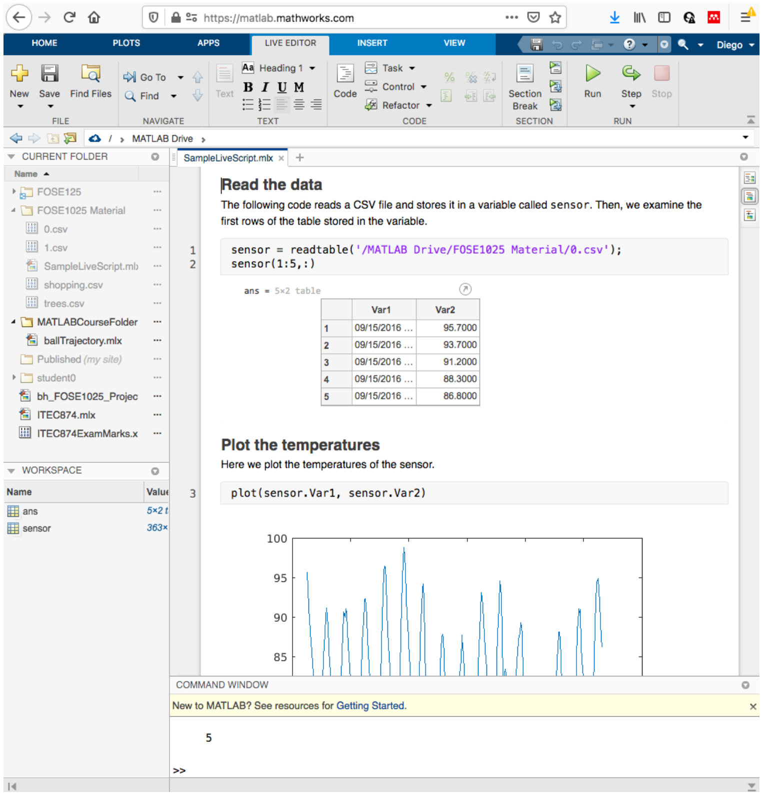
We can experiment with the data to see if any of the variables in the table are correlated with highway fatalities. It appears that highway fatality rates are lower in states with a higher percentage urban population.

```
varName = 'urbanPopulation';
scatter(fatalities.(varName),rate)
xlabel(varName)
ylabel('Percent Fatalities per Million Vehicle Miles')

hold on
xmin = min(fatalities.(varName));
xmax = max(fatalities.(varName));
p = polyfit(fatalities.(varName),rate,1);
plot([xmin xmax], polyval(p,[xmin xmax]))
```

Demonstration of a Live Script

SampleLiveScript.mlx



Read the data

The following code reads a CSV file and stores it in a variable called sensor. Then, we examine the first rows of the table stored in the variable.

```
sensor = readtable('/MATLAB Drive/FOSE1025 Material/0.csv');
sensor(1:5,:)
```

ans = 5×2 table

	Var1	Var2
1	09/15/2016 ...	95.7000
2	09/15/2016 ...	93.7000
3	09/15/2016 ...	91.2000
4	09/15/2016 ...	88.3000
5	09/15/2016 ...	86.8000

Plot the temperatures

Here we plot the temperatures of the sensor.

```
plot(sensor.Var1, sensor.Var2)
```

```
   5

>>
```

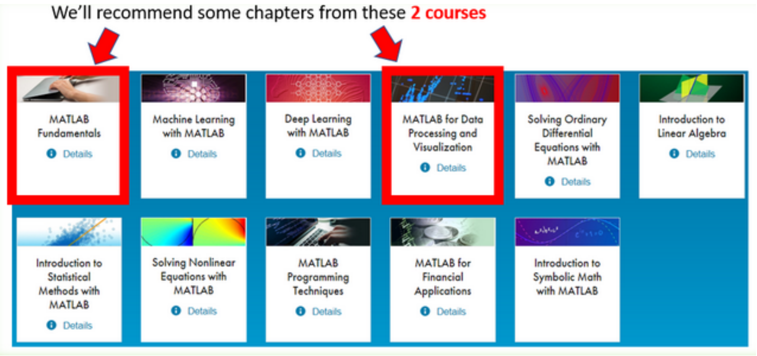
3 MATLAB

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/*

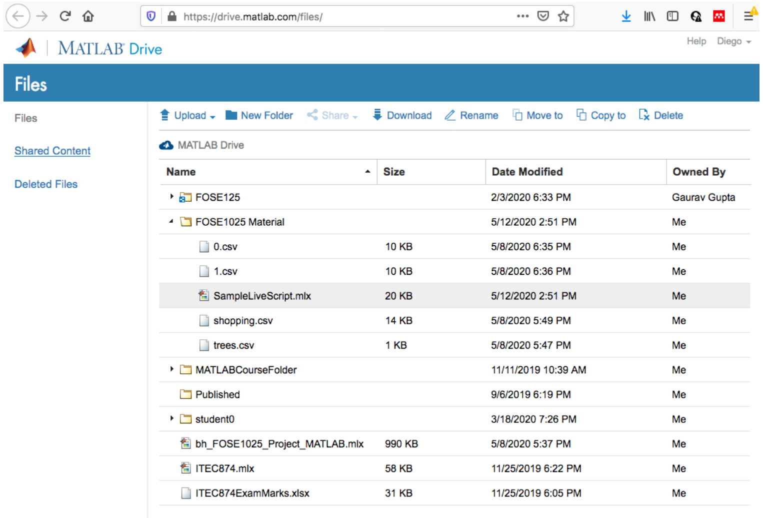


We'll recommend some chapters from these 2 courses

MATLAB Fundamentals
Details
Machine Learning with MATLAB
Details
Deep Learning with MATLAB
Details
MATLAB for Data Processing and Visualization
Details
Solving Ordinary Differential Equations with MATLAB
Details
Introduction to Linear Algebra
Details
Introduction to Statistical Methods with MATLAB
Details
Solving Nonlinear Equations with MATLAB
Details
MATLAB Programming Techniques
Details
MATLAB for Financial Applications
Details
Introduction to Symbolic Math with MATLAB

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.



Name	Size	Date Modified	Owned By
FOSE125		2/3/2020 6:33 PM	Gaurav Gupta
FOSE1025 Material		5/12/2020 2:51 PM	Me
0.csv	10 KB	5/8/2020 6:35 PM	Me
1.csv	10 KB	5/8/2020 6:36 PM	Me
SampleLiveScript.mlx	20 KB	5/12/2020 2:51 PM	Me
shopping.csv	14 KB	5/8/2020 5:49 PM	Me
trees.csv	1 KB	5/8/2020 5:47 PM	Me
MATLABCourseFolder		11/11/2019 10:39 AM	Me
Published		9/6/2019 6:19 PM	Me
student0		3/18/2020 7:26 PM	Me
bh_FOSE1025_Project_MATLAB.mlx	990 KB	5/8/2020 5:37 PM	Me
ITEC874.mlx	58 KB	11/25/2019 6:22 PM	Me
ITEC874ExamMarks.xlsx	31 KB	11/25/2019 6:05 PM	Me

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:

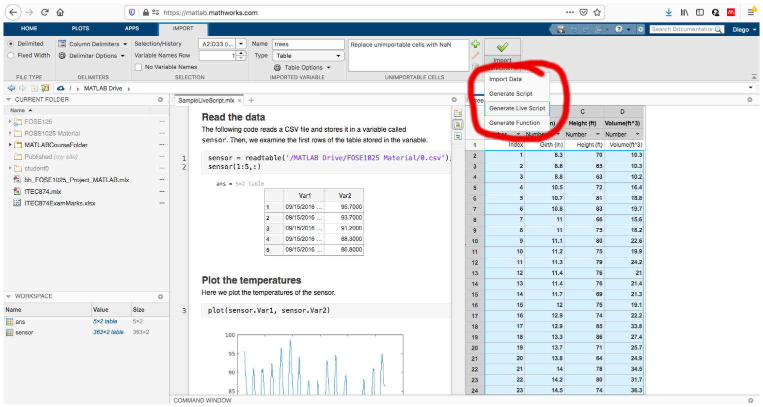
```
table0 = readtable(”0.csv”);
table1 = readtable(”1.csv”);
table = [table0; table1];
```

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.



● Delimited ○ Fixed Width

Selection/History: A2:D33 (i...

Variable Names Row: 1

Name: trees

Type: Table

Replace unimportable cells with NaN

WORKSPACE

Name	Value	Size
ans	5×2 table	5×2
sensor	363×2 table	363×2

Read the data

The following code reads a CSV file and stores it in a variable called `sensor`. Then, we examine the first rows of the table stored in the variable.

```
sensor = readtable('/MATLAB Drive/FOSE1025 Material/0.csv');
sensor(1:5,:)
```

```
ans = 5×2 table
```

	Var1	Var2
1	09/15/2016 ...	95.7000
2	09/15/2016 ...	93.7000
3	09/15/2016 ...	91.2000
4	09/15/2016 ...	88.3000
5	09/15/2016 ...	86.8000

Plot the temperatures

Here we plot the temperatures of the sensor.

```
plot(sensor.Var1, sensor.Var2)
```

			C	D
		n)	Height (ft)	Volume(ft^3)
		Number	Number	Number
1	Index	Girth (in)	Height (ft)	Volume(ft^3)
2	1	8.3	70	10.3
3	2	8.6	65	10.3
4	3	8.8	63	10.2
5	4	10.5	72	16.4
6	5	10.7	81	18.8
7	6	10.8	83	19.7
8	7	11	66	15.6
9	8	11	75	18.2
10	9	11.1	80	22.6
11	10	11.2	75	19.9
12	11	11.3	79	24.2
13	12	11.4	76	21
14	13	11.4	76	21.4
15	14	11.7	69	21.3
16	15	12	75	19.1
17	16	12.9	74	22.2
18	17	12.9	85	33.8
19	18	13.3	86	27.4
20	19	13.7	71	25.7
21	20	13.8	64	24.9
22	21	14	78	34.5
23	22	14.2	80	31.7
24	23	14.5	74	36.3

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