FOSE1025 — Scientific Computing

Week 6 Lecture 1: Towards Using Scripts for Reproducibility

Diego Mollá

Department of Computer Science
Macquarie University

FOSE1025 2023H1

Programme

1. Excel and MATLAB for Science
2. Scripts for Reproducibility
3. MATLAB

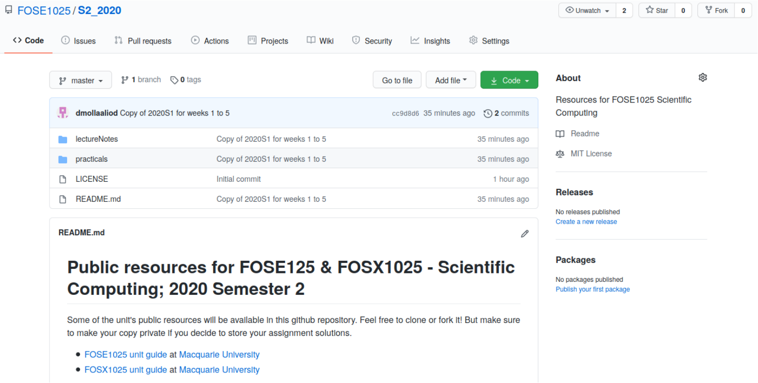
Reading

- Lecture notes
- Readings listed in iLearn — Week 6

Announcements

1. Lecture notes in Echo360.
2. Material in github (next slide).

FOSE1025's public github page



FOSE1025 / **S2_2020**

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

About

Resources for FOSE1025 Scientific Computing

Readme

MIT License

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

README.md

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

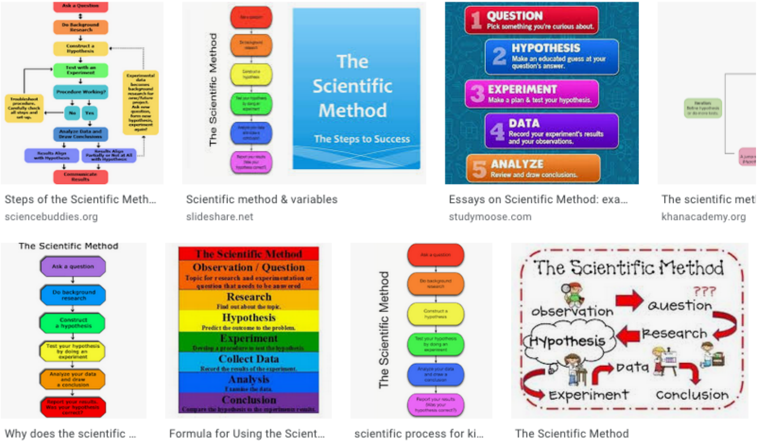
Some of the unit's public resources will be available in this github repository. Feel free to clone or fork it! But make sure to make your copy private if you decide to store your assignment solutions.

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

Programme

1. Excel and MATLAB for Science
2. Scripts for Reproducibility
3. MATLAB

The Scientific Method


Ask a Question
Do Background Research
Construct a Hypothesis
Test with an Experiment
Procedure Working?
No
Yes
Analyze Data and Draw Conclusions
Communicate Results
The Scientific Method
The Scientific Method
The Steps to Success
1 QUESTION
Pick something you're curious about.
2 HYPOTHESIS
Make an educated guess at your question's answer.
3 EXPERIMENT
Make a plan & test your hypothesis.
4 DATA
Record your experiment's results and your observations.
5 ANALYZE
Review and draw conclusions.
The Scientific Method
Ask a question
Do background research
Construct a hypothesis
Test your hypothesis by doing an experiment
Analyze your data and draw a conclusion
Report your results. Was your hypothesis correct?
The Scientific Method
Observation / Question
Topic for research and experimentation or question that needs to be answered
Research
Find out about the topic.
Hypothesis
Predict the outcome to the problem.
Experiment
Develop a procedure to test the hypothesis.
Collect Data
Record the results of the experiment.
Analysis
Examine the data.
Conclusion
Compare the hypothesis to the experiments results.
The Scientific Method
The Scientific Method
observation
???
question
Research
Hypothesis
Data
Experiment
conclusion

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 and MATLAB to Manage Data in Science

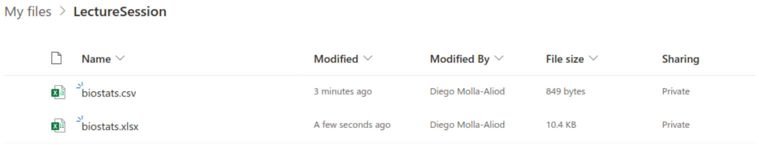
We are covering these aspects in FOSE1025:

- Represent data in Excel — Week 2.
- Represent data in MATLAB — Week 5.
- Explore data in Excel — Weeks 3 & 4.
- Visualise data in Excel — Week 5.
- (you are here)
- 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.
- Analyse, summarise, interpret the data (MATLAB) — Week 9.
- Ethics of Data — Week 10.

Importing Data

Excel Files

- Excel saves files in a special format.
- The name of these files ends with .xlsx
- Other programs (e.g. MATLAB) can read Excel files.
- But there are many other formats!



My files > LectureSession

Name	Modified	Modified By	File size	Sharing
biostats.csv	3 minutes ago	Diego Molla-Aliod	849 bytes	Private
biostats.xlsx	A few seconds ago	Diego Molla-Aliod	10.4 KB	Private

Importing Data

Comma Separated Values

CSV — Comma Separated Values

- CSV is a very simple file format used by many applications.
- Each line represents a table row.
- Different values in the row are separated with a comma.
  - We say that comma is the delimiter.
  - Other common delimiters are: tabulator space (tab), semicolon (;).
  - In some file formats, the data fields are determined by setting a fixed text width to each field.

Example of a CSV File

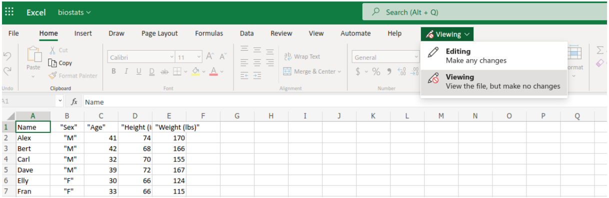
biostats.csv from
https://people.sc.fsu.edu/˜jburkardt/data/csv/csv.html

```
"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
```

Importing a CSV File into Excel Online

The easy option: upload & click

1. Upload the file to OneDrive.
2. Click on the file.
   - Excel Online will open the file in viewing mode.
3. Set the viewing mode to Editing.
   - Excel Online will create a copy of the file and save it as an Excel (.xlsx) file.



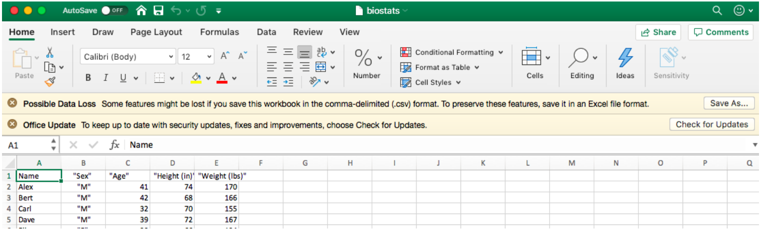
	A	B	C	D	E	F
1	Name	"Sex"	"Age"	"Height (i	"Weight (lbs)"	
2	Alex	"M"	41	74	170	
3	Bert	"M"	42	68	166	
4	Carl	"M"	32	70	155	
5	Dave	"M"	39	72	167	
6	Elly	"F"	30	66	124	
7	Fran	"F"	33	66	115	

Importing a CSV File into Excel Online

Option 2: copy & paste

1. Double click on the file.
    - The file will open in your desktop.
    - If you have Excel installed, Excel will open the file.
2. Create a blank Excel workbook in Excel Online.
3. Copy and paste from the desktop application to Excel Online.
4. Save or rename the file.

Careful if you double-click on a CSV file in your desktop!



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

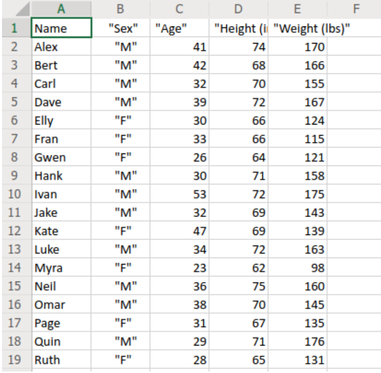
A1 fx Name

	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 (more on this in future lectures).



	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?

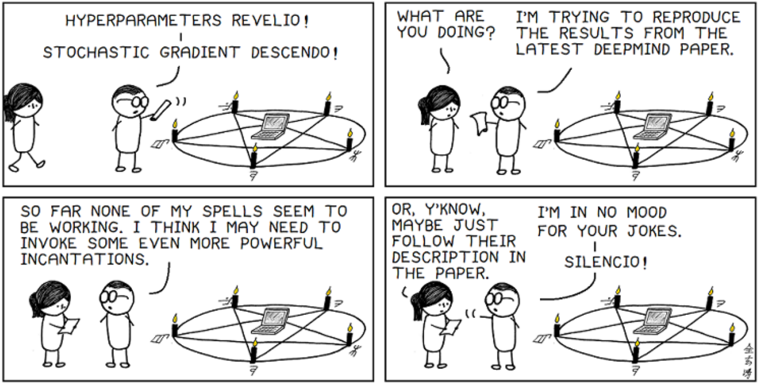
Programme

1. Excel and MATLAB for Science
2. Scripts for Reproducibility
3. MATLAB

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: Running Scripts in MATLAB Online

- In this demonstration, we will run MATLAB code 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 using your student email address.
- Do not use your student password (create a new password).

File vectorsMatrices.m

```
data1 = zeros(1, 5) % row vector
data2 = zeros(5, 1) % column vector
mult = 5.4
data3 = ones(10, 1) * mult %scalar multiplication
data4 = data3 + 5 %scalar addition
taxicab1 = [10 70 20 90]
taxicab2 = taxicab1' %converting between row and column vector
myFirstMatrix1 = [10 70; 20 90; 30 80]
myFirstMatrix2 = myFirstMatrix1 + 2.5 %scalar addition

col1 = [10 70 20 90 30 80]'
col2 = [40 80 20 60 30 10]'
col3 = [20 10 0 -100 -2000 0]'
sumCols = col1 + col2 + col3
%you can add or subtract vector/matrices of the same size

mat1 = [10 70 20; 30 90 80]
mat2 = [50 10 90; 100 30 70]
mat3 = (mat1 + mat2)'

mat3(3,1) %matrix(row number, column number)
```

Running File vectorsMatrices.m in MATLAB Online

In this demonstration we will:

1. Upload a MATLAB script to MATLAB Drive.
2. Open the MATLAB script.
3. Run the MATLAB script.

MATLAB File Format

- The file extension .m indicates it is a MATLAB script.
- If you open the file using a text file, you will see that it contains plain text.
- If you double click on the file and MATLAB is installed in your computer, MATLAB will open the file.

Scripting Languages and Reproducibility

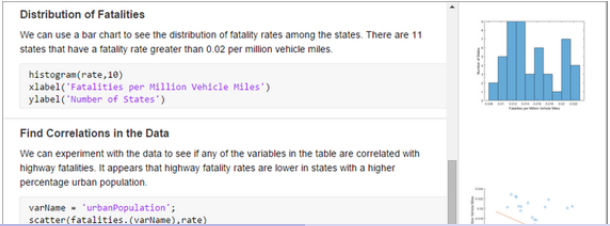
The Problem with Regular Scripts

- Regular scripts (e.g. MATLAB's .m files) are good if we want to run code.
- But what if we can to keep record of a scientific experiment?
- We will want to supplement the code with comments and explanations.
- We will also want to show the output of some of the code, e.g. plots.

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')
```

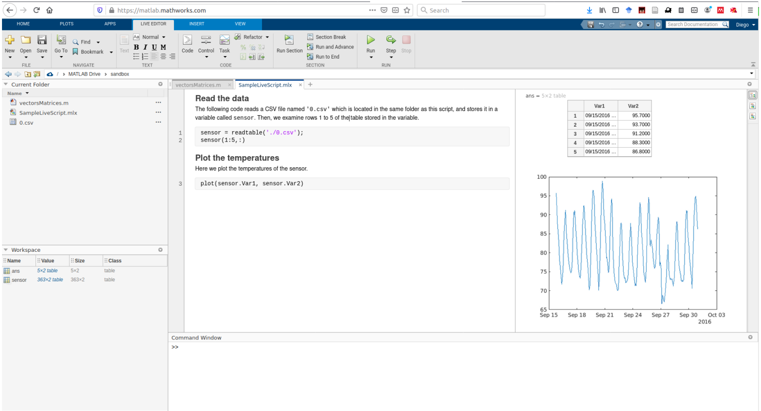
Find Correlations in the Data

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)
```

Demonstration of a Live Script

File: SampleLiveScript.mlx



Read the data

The following code reads a CSV file named `'0.csv'` which is located in the same folder as this script, and stores it in a variable called `sensor`. Then, we examine rows 1 to 5 of the table stored in the variable.

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

Plot the temperatures

Here we plot the temperatures of the sensor.

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

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

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

Running SampleLiveScript.mlx in MATLAB Online

In this demonstration we will:

1. Upload a MATLAB Live Script to MATLAB Drive.
2. Open the MATLAB Live Script by double-clicking on the uploaded file.
3. Run the MATLAB Live Script.

MATLAB Live Script Format

- MATLAB Live Scripts use the file extension .mlx.
- This is a special file format. If you open the file with a text editor, you will see garbage.
- If you double click on the file and MATLAB is installed in your computer, MATLAB will open the file.

Programme

1. Excel and MATLAB for Science
2. Scripts for Reproducibility
3. MATLAB

Why MATLAB?

- MATLAB is a scripting language.
- Includes types designed to store and manipulate data.
  - Vectors and 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.
- Other programming languages have attempted to integrate some of MATLAB's features.
  - Matrices, tables
  - Plots
  - Interactive notebooks

Accessing MATLAB and MATLAB Online

Accessing MATLAB

- Macquarie University has a license for students: https://au.mathworks.com/academia/tah-portal/macquarie-university-916052.html
- MATLAB Online: https://matlab.mathworks.com/

Learning MATLAB

- Getting started: https://au.mathworks.com/help/matlab/getting-started-with-matlab.html
- Self-paced courses: https://matlabacademy.mathworks.com/

Importing CSV Files in MATLAB

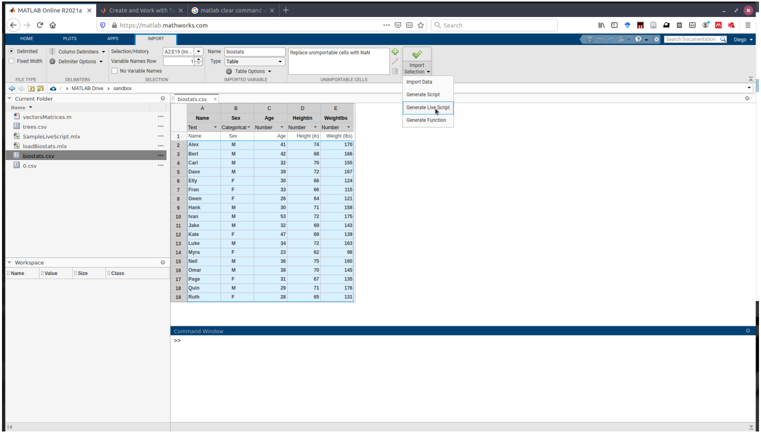
- MATLAB Fundamentals, Chapter 10, “Tables of Data”
- https://au.mathworks.com/help/releases/R2022a/matlab/matlab_prog/create-a-table.html
- MATLAB can store tables into variables.
- You can use the MATLAB “Import” wizard to load CSV files (next slide).
- Or you can use the readtable instruction.
  - trees = readtable(" biostats.csv");

  (The "Import" tool will generate a MATLAB script that ultimately executes readtable(" biostats.csv", opts), where opts specifies options that override readtable's defaults.)

Generate a live script that imports file biostats.csv

1. Upload file biostats.csv to MATLAB Drive.
2. Double-click on biostats.csv. The “Import” tool will open.
3. Select the correct options (the defaults will not always work):
   - Name of the MATLAB variable where the table will be stored.
   - Names of the columns.
   - Data types of the columns.
   - The range of the data to import.
4. Select the correct import option in the ”Import Selection” dropbox.
   - Select ”Generate Live Script”
5. Save the live script and execute it (click on ”Run”).
   - The live script will be saved into a file with extension .mlx
   - This file is saved in MATLAB Drive “in the cloud”, not in your desktop computer.

Generating the live script



● Delimited
○ Fixed Width

Selection/History: A2:E19 (ini...

Variable Names Row: 1

□ No Variable Names

Name: biostats

Type: Table

Replace unimportable cells with NaN

Current Folder:

- vectorsMatrices.m
- trees.csv
- SampleLiveScript.mlx
- loadBiostats.mlx
- biostats.csv
- 0.csv

Workspace:

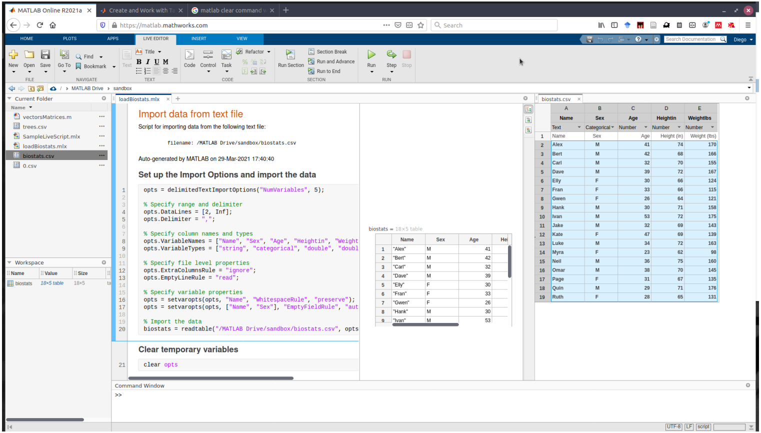
Name	Value	Size	Class

biostats.csv

	A	B	C	D	E
	Name	Sex	Age	Heightin	Weightlbs
	Text	Categorical	Number	Number	Number
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
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

```
>>
```

After running the live script



Import data from text file

Script for importing data from the following text file:

```
filename: /MATLAB Drive/sandbox/biostats.csv
```

Auto-generated by MATLAB on 29-Mar-2021 17:40:40

Set up the Import Options and import the data

```
opts = delimitedTextImportOptions("NumVariables", 5);

% Specify range and delimiter
opts.DataLines = [2, Inf];
opts.Delimiter = ",";

% Specify column names and types
opts.VariableNames = ["Name", "Sex", "Age", "Heightin", "Weight
opts.VariableTypes = ["string", "categorical", "double", "doubl

% Specify file level properties
opts.ExtraColumnsRule = "ignore";
opts.EmptyLineRule = "read";

% Specify variable properties
opts = setvaropts(opts, "Name", "WhitespaceRule", "preserve");
opts = setvaropts(opts, ["Name", "Sex"], "EmptyFieldRule", "aut

% Import the data
biostats = readtable("/MATLAB Drive/sandbox/biostats.csv", opts
```

Clear temporary variables

```
clear opts
```

biostats = 18×5 table

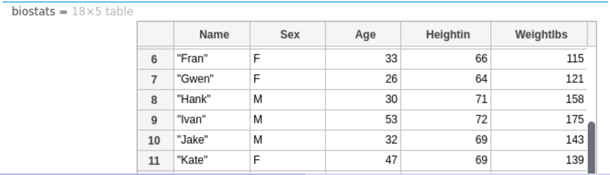
	Name	Sex	Age	Hei
1	"Alex"	M	41	
2	"Bert"	M	42	
3	"Carl"	M	32	
4	"Dave"	M	39	
5	"Elly"	F	30	
6	"Fran"	F	33	
7	"Gwen"	F	26	
8	"Hank"	M	30	
9	"Ivan"	M	53	

biostats.csv

	A	B	C	D	E
	Name	Sex	Age	Heightin	Weightlbs
	Text	Categorical	Number	Number	Number
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
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

Anatomy of a MATLAB Table

- A MATLAB table has rows and columns.
- All cells in the column are of the same data type.
  - For example, the "Name" column of the "biostats" table has cells of type string.
- Cells in different columns of the same table may have different data types.
- When you load a CSV file and store it into a MATLAB variable, you can determine the data type associated with each column.
  - MATLAB will guess the best data type for each column, but sometimes you want to override these guesses.



biostats = 18×5 table

	Name	Sex	Age	Heightin	Weightlbs
6	"Fran"	F	33	66	115
7	"Gwen"	F	26	64	121
8	"Hank"	M	30	71	158
9	"Ivan"	M	53	72	175
10	"Jake"	M	32	69	143
11	"Kate"	F	47	69	139

Accessing and Modiying Data in MATLAB Tables

We will see how we can do the following in MATLAB:

- Accessing all the values of a column.
- Accessing some values of a column.
- Accessing all the columns of a row.
- Accessing some of the columns of a row.
- Modifying the data of the table.

Accesing data in MATLAB tables

Example: biostats.csv

https://au.mathworks.com/help/releases/R2022a/matlab/matlab_prog/access-data-in-a-table.html

- Accessing all values of a column: `names = biostats.Name`
  - This will create a column vector and store it in variable names.
- Accessing some values of a column: `names1to5 = biostats.Name(1:5)`
  - This will create a column vector with the first 5 values of the Name column and store it in variable names1to5.
- Accessing the second row:
  - `secondrow1 = biostats(2,:)` as table
  - `secondrow2 = biostats{2,4:5}` as an array
  - `secondrow3 = biostats{2,1:5}` will not work because the result cannot be an array! (there are multiple data types)
- Accessing columns: `heightcm = biostats.Heightin * 2.54`

Modifying data in MATLAB tables

Example: biostats.csv

https://au.mathworks.com/help/releases/R2022a/matlab/matlab_prog/access-data-in-a-table.html

- Creating a new column: `biostats.Heightcm = heightcm`
  - This will create a new column in table `biostats`. The name of the column is Heightcm. The values of variable heightcm will be copied (assigned to) this new column.
- Modifying part of a column:
  `biostats.Heightcm(1:3) = [0;0;0]` (Q: why the semicolons?)
  - This will set the first 3 elements of column Heightcm to zero. Note that we must assign a column vector.
- Modifying a row. The following two are equivalent:
  - `biostats{7,["Age" "Weight_lbs]} = [31 110]`: indexing by column name.
  - `biostats{7,[3 5]} = [31 110]`: indexing by column number.

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

Assessments this week

- In-class test 2 during your SGTA class.
    - This is not a hurdle but worth 25% of the unit assessment.
- Professiional hurdle (quiz): Submit by Friday 31 March, 11:55pm.

Next week

- Week 7 lecture: Cleaning data.
- No assessments are due on week 7.
- 2 weeks without class after week 7.