## Data Files - Spreadsheets & File Types

Suzanne Little, [suzanne.little@dcu.ie](mailto:suzanne.little@dcu.ie)

### Exploring Spreadsheets

Spreadsheet basics: Terminology (from<https://zapier.com/learn/google-sheets/google-sheets-tutorial/>)

* **Cell**: A single data point or element in a spreadsheet.
* **Column**: A vertical set of cells.
* **Row**: A horizontal set of cells.
* **Range**: A selection of cells extending across a row, column, or both.
* **Function**: A built-in operation from the spreadsheet app, which can be used to calculate cell, row, column, or range values, manipulate data, and more.
* **Formula**: The combination of functions, cells, rows, columns, and ranges used to obtain a specific result → Like a mini-program!
* **Worksheet (Sheet)**: The named sets of rows and columns making up your spreadsheet; one spreadsheet can have multiple sheets
* **Spreadsheet**: The entire document containing your worksheets

**Some references if you want to brush up on your spreadsheet skills**

* Chapter 1: "Everything You Ever Needed to Know about Spreadsheets but Were Too Afraid to Ask" from "Data smart: using data science to transform information into insight", John W. Foreman (2018)<https://dcu.primo.exlibrisgroup.com/permalink/353DCU_INST/jrp0g3/alma991005443517807206>
* Part 1 of “Excel 2013 for Scientists”, Gerard Verschuuren (2014) <https://dcu.primo.exlibrisgroup.com/permalink/353DCU_INST/jrp0g3/alma991005460440507206>
* See also Introduction to Spreadsheets in datacamp - <https://app.datacamp.com/learn/courses/introduction-to-spreadsheets>

### 

### Exercise: Complete the following steps:

1. Download [DataSampler.csv](https://raw.githubusercontent.com/suzannelittle/ca682i/master/data/sampler/DataSampler.csv).
2. Open the file in Notepad or a similar programme and look at the CSV format. What is different about the 8th line?
3. Open the file in a spreadsheet programme (Excel or Google Sheets) and explore the data.
4. What are the *types* of the data in each column? Refer to your notes for some terms you can use.
5. Use the spreadsheet functions to calculate values like mean, standard deviation for the numerical columns. For which columns do these statistical descriptors make sense?
6. [Advanced] Try creating histograms for columns C & D (Process is something like: select all values in the column then Insert → Statistics Chart → Histogram). Can you see differences?

### 

### Exercise: Common data science file types

1. CSV (& TSV & TXT)
2. JSON
3. XLS or XLSX
4. SQL
5. PDF
6. HTML
7. DOC or DOCX
8. HDF5
9. ZIP (or GZ or TGZ)
10. XML
11. Can you identify the main program associated with each? (That is, the programme that normally creates the file)
12. Which of these file types do you know how to read into python? (or any other programming language if you’re not a python programmer)
13. For each file type: is it a binary or a text file?