Lab 2: Correlating Data Sources

The objective of this lab is to practice:

- Obtaining data from web resources.
- Using text transformations on data to put it into a form suitable for a spreadsheet.
- Making scatter plots.
- Computing correlations.

This will require the use of Atom that you can download here:
https://atom.io/ click download (for windows users)

or

https://github.com/atom/atom/releases/tag/v1.54.0 (look for atom-windows.zip or atom-mac.zip extract the files and start atom)

You can use Notepad++ or any other text editor that allows the use of regular expressions but the examples that follow will be using the Atom editor which is available on Windows, Mac and Linux.

Some resources if you get stuck on a step during the lab:

- Excel SUM function
- Excel CORREL function

1. Cambridge Colleges

The University of Cambridge is made up of a central university and a collection of colleges. The colleges are responsible for admitting students to the university. Colleges are the center of Cambridge life—students live, socialize and do most of their studying in their college. The 31 Cambridge colleges were founded between 1284 (Peterhouse) and 1977 (Robinson).

The Cambridge colleges hold substantial collections of wine. There has recently been a flurry of stories about how much wine the Cambridge colleges' purchase. One article in *The Telegraph* reports that the university's colleges spent £3,000,000 in one year alone. [1]. **We will be exploring a hypothetical scenario using different figures.**

2. Collect the Wine Data

The file "Lab2Data.txt" contains **fictitious** data about how much each college spent. **The figures used in this file are not the actual figures spent by the colleges**. Create a new file in the atom editor by clicking file->New File. Copy the contents of "Lab2Data.txt" into the new file in the atom editor. We now need to edit this data into a tab-separated value (TSV) format that can be pasted directly into Excel. Since there

are only 31 colleges, you could edit the lines each by hand into the form required, but if there were 31,000 lines, you would want to do it a smarter way.

Do the following:

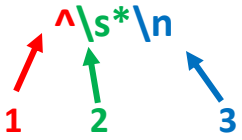
1. **Remove the blank lines from the file.**
   First save the new file you created with the data you copied from Lab2Data.txt. Save the file as goodwine.txt.

   Next select the Find menu and click on Replace in Buffer. In the 'Find in current buffer" box, put the regular expression below.

   ^\s\n *(atom users)*
   ^\s*\r\n *(Notepad++ Users Only)*

   At the bottom right of the window you will see a ".*" and a "Aa" button. Select both to enable *Regular expressions and also enable case sensitivity(just to make sure no issues occur)*.
   Then click on the Replace all button.

   If done correctly, this regular expression should remove all blank lines. The explanation for the pattern is as follows:



   1. Match the beginning of a line.
   2. Match **zero or more** white space characters (can be spaces, tabs, etc.)
   3. Match a line break. Note that **\n** represents the invisible characters that make up a line break.

2. **Fix up the text.**
   Correct "Emma" to be "Emmanuel".
   Change "Caius" to be "Gonville and Caius"

3. **Remove pound symbols.** Remove the pound signs (£) from all of the amounts. You can accomplish this by selecting the "Find" menu and choosing "Replace in Buffer". In the "Find in current buffer" box, put "£" and make sure the "Replace in current buffer" box is empty and contains no spaces. Click "Replace All". Note that on Windows you can type a £ symbol by holding down the Alt key and pressing 0163 on the numerical keypad or just copy it from the text in the file itself.

4. **Add tabs.** Convert the text ": " (colon space) to a tab. Do this by selecting the "Find" menu and choosing "Replace in Buffer". In the "Find in current buffer" box, put ": " (a colon followed by a space). In the "Replace in current buffer" box, put "\t". Click "Replace All".

5. **Paste into a spreadsheet.** Open Excel. Copy all the text from the editor. You can do this by typing Control-A (select all) followed by Control-C (copy) in the editor window. In a blank Excel workbook, highlight cell **A2** and press Control-V (paste) to paste the data into Excel.

6. **Format the spreadsheet.** Format the amounts as currency (make sure the pound sign (£) is selected as the symbol). Give each column an appropriate header and ensure the columns are wide enough to show all data (i.e. there are no cells with ##### in them).

3. Collect More Data About the Colleges From Wikipedia

NOTE: For this step ensure that you use Firefox as your browser as Internet Explorer can cause some issues when copying and pasting the data.

1. The Wikipedia article "Colleges of the University of Cambridge" [3] has data about the colleges, including founding dates, enrollments and the value of fixed assets. Copy and paste the table from this web page into Atom.

2. In addition to removing blank lines, use regular expression to remove the citations, numbers between two square brackets ([]s). Recall from the assigned reading that we can match the literal '[' and ']' characters by escaping them with a backslash (e.g. \[and \]). We can match a single digit in a few ways such as \d or [0-9].

3. Manually make sure the column headers are on one line in the text file. Also remove the totals on the last line if you copied them in.

4. Make sure you also remember to remove pound signs (£) like we did in step 2 and remove the "N/A"s in the "Net Assets" and "Assets per student" columns.

5. Once TSV formatted, highlight cell C1 in your Excel sheet and paste the data. Remove the scarf colours, Abb, website, and notes columns from the table. Clean up the founding years so that

each entry is just a single number (pick the earlier year when there are two). Ensure the college name columns match (same college name on each row in both columns) and remove the extra college name column.

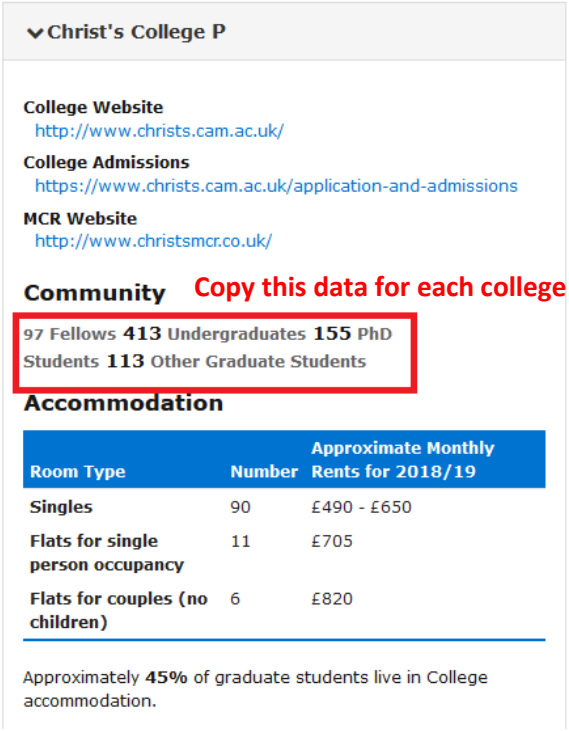
6. Format the data appropriately and ensure the correct currency type is used for endowment, net assets and assets per student.

7. Your Excel sheet should currently look something similar to this (note that values might be slightly different if the Wikipedia page has recently been updated):

	A	B	C	D	E	F	G	H	I	J	K
1	**College**	**Wine Spending**	**Founded**	**U**	**P**	**Male %**	**Female %**	**Total**	**Endowment**	**Net Assets (Mostly 2014)**	**Assets per student**
2	Christ's	£71,055.00	1505	450	91	58	42	541	£142,177,000.00	£149,700,000.00	£276,710.00
3	Churchill	£87,685.00	1960	476	228	71	29	704	£34,750,000.00	£143,795,000.00	£204,254.00
4	Clare	£79,989.00	1326	473	182	52	48	655	£106,400,000.00	£138,494,000.00	£211,441.00
5	Clare Hall	£17,400.00	1966	0	155	47	53	155	£14,851,457.00	£14,219,450.00	£91,738.00
6	Corpus Christi	£79,254.00	1352	266	201	60	40	467	£97,363,000.00	£209,187,829.00	£455,747.00
7	Darwin	£17,510.00	1964	0	674	54	46	674	£23,732,104.00	£47,629,759.00	£70,677.00
8	Downing	£77,798.00	1800	440	183	66	34	623	£40,064,267.00	£149,404,000.00	£239,814.00
9	Emmanuel	£131,127.00	1584	470	280	51	49	750	£237,003,395.00	£253,729,224.00	£338,305.00
10	Fitzwilliam	£23,028.00	1869	502	186	63	37	688	£58,537,000.00	£79,124,000.00	£115,006.00
11	Girton	£30,051.00	1869	455	321	53	47	776	£44,662,000.00	£144,451,000.00	£186,148.00
12	Gonville and Caius	£96,994.00	1348	546	173	60	40	719	£46,777,421.00	£185,343,378.00	£257,779.00
13	Homerton	£27,974.55	1768	593	588	37	63	1181	£145,981,706.00	£138,417,093.00	£117,203.00
14	Hughes Hall	£14,033.58	1885	60	500	61	39	560	£3,224,988.00		
15	Jesus	£212,256.00	1496	503	201	57	43	704	£161,099,438.00	£276,509,040.00	£392,769.00
16	King's	£338,559.00	1441	394	187	57	43	581	£109,736,000.00	£179,656,000.00	£309,219.00
17	Lucy Cavendish		1965	110	110	0	100	220	£13,253,000.00	£23,825,000.00	£108,295.00
18	Magdalene	£68,192.00	1428	366	127	54	46	493	£46,783,512.00	£98,702,257.00	£200,207.00
19	Murray Edwards	£32,917.00	1954	387	55	0	100	442	£32,537,871.00	£74,812,930.00	£169,260.00
20	Newnham	£27,263.00	1871	412	112	0	100	524	£49,543,320.00	£156,451,793.00	£298,572.00
21	Pembroke	£141,692.00	1347	442	155	53	47	597	£69,605,878.00	£135,332,773.00	£226,688.00
22	Peterhouse	£82,133.00	1284	260	110	57	43	370	£268,059,000.00	£271,052,000.00	£765,684.00
23	Queens'	£111,112.64	1448	535	297	57	43	832	£69,000,000.00	£90,204,100.00	£108,418.00
24	Robinson	£44,722.39	1977	422	73	60	40	495	£77,542,000.00	£82,112,000.00	£165,883.00
25	St Catharine's	£62,432.00	1473	462	159	52	48	621	£62,200,000.00	£97,395,000.00	£156,836.00
26	St Edmund's	£19,304.00	1896	120	379	69	31	499	£15,212,000.00	£23,733,000.00	£78,052.00
27	St John's	£260,064.00	1511	588	243	59	41	831	£691,500,000.00	£700,545,000.00	£843,014.00
28	Selwyn	£49,498.00	1882	361	269	55	45	630	£75,957,306.00	£94,953,363.00	£183,308.00
29	Sidney	£97,507.00	1596	371	135	63	37	506	£36,212,762.00	£110,295,436.00	£217,975.00
30	Trinity	£223,291.98	1546	671	359	63	37	1030	£1,008,016,030.00	£1,028,272,000.00	£998,322.00
31	Trinity Hall	£127,186.00	1350	384	196	54	46	580	£47,896,088.00	£223,677,179.00	£385,650.00
32	Wolfson	£39,647.10	1965	119	385	64	36	504	£31,533,000.00	£33,171,000.00	£65,815.00

4. Collect More Data from The Cambridge Website

The Cambridge administration provides data on the number of Fellows at each college. These are typically faculty members who have dining rights and can walk on the grass [4]. Collect this data from the graduate admissions website by clicking on each college name and pasting the data under "Community" into a new line in Notepad++/Sublime Text 3. **Ensure that the data for each college is in the same order as the colleges in your Excel sheet** (Note that the colleges starting with "St." are in a different order on the webpage).



Christ's College P

College Website
http://www.christs.cam.ac.uk/

College Admissions
https://www.christs.cam.ac.uk/application-and-admissions

MCR Website
http://www.christsmcr.co.uk/

Community **Copy this data for each college**

97 Fellows **413** Undergraduates **155** PhD Students **113** Other Graduate Students

Accommodation

Room Type	Number	Approximate Monthly Rents for 2018/19
Singles	90	£490 - £650
Flats for single person occupancy	11	£705
Flats for couples (no children)	6	£820

Approximately **45%** of graduate students live in College accommodation.

When finished, your *Atom* document should look like this:

```
97 Fellows 413 Undergraduates 142 PhD Students 82 Other Graduate Students
175 Fellows 470 Undergraduates 245 PhD Students 99 Other Graduate Students
120 Fellows 504 Undergraduates 120 PhD Students 180 Other Graduate Students
67 Fellows 0 Undergraduates 96 PhD Students 85 Other Graduate Students
59 Fellows 280 Undergraduates 160 PhD Students 80 Other Graduate Students
65 Fellows 0 Undergraduates 383 PhD Students 273 Other Graduate Students
65 Fellows 404 Undergraduates 147 PhD Students 131 Other Graduate Students
87 Fellows 430 Undergraduates 121 PhD Students 99 Other Graduate Students
56 Fellows 492 Undergraduates 200 PhD Students 170 Other Graduate Students
128 Fellows 500 Undergraduates 135 PhD Students 82 Other Graduate Students
114 Fellows 540 Undergraduates 280 Graduate Students
47 Fellows 629 Undergraduates 174 PhD Students 414 Other Graduate Students
59 Fellows 100 Undergraduates 200 PhD Students 300 Other Graduate Students
100 Fellows 500 Undergraduates 280 PhD Students 220 Other Graduate Students
128 Fellows 399 Undergraduates 217 PhD Students 84 Other Graduate Students
40 Fellows 144 Undergraduates 130 PhD Students 130 Other Graduate Students
125 Fellows 345 Undergraduates 151 NOTAFS/PhD Students 62 Other Graduate Students.
58 Fellows 360 Undergraduates 81 PhD Students 69 Other Graduate Students
45 Fellows 363 Undergraduates 150 Annual Graduate Intake
102 Fellows 444 Undergraduates 170 PhD Students 105 Other Graduate Students
42 Fellows 280 Undergraduates 77 Annual Graduate Intake
84 Fellows 511 Undergraduates 258 PhD Students 232 Other Graduate Students
82 Fellows 386 Undergraduates 51 PhD Students 121 Other Graduate Students
60 Fellows 430 Undergraduates
117 Fellows 117 Mature Undergraduates 154 PhD Students 232 Other Graduate Students
159 Fellows 597 Undergraduates 230 PhD Students 92 Other Graduate Students
58 Fellows 377 Undergraduates 130 PhD Students 92 Other Graduate Students
67 Fellows 348 Undergraduates 144 PhD Students 130 Other Graduate Students
185 Fellows 700 Undergraduates 350 Graduate Students
75 Fellows 380 Undergraduates 163 PhD Students 90 Other Graduate Students
271 Fellows 135 Undergraduates 340 PhD Students 280 Other Graduate Students
```

To create a tab separated document to paste into Excel, we need to replace the words between the numbers with tabs. To accomplish this, click on the “Find” menu and choosing “Replace in Buffer”. Ensure the “.*” (Regular Expression) button is selected. This allows us to use regular expression to replace matched text in the document. Enter the following in “Find in current buffer”

```
[a-zA-Z ./]+
```

Note the space between the “Z” and the “]”. Enter “\t” (without quotes) into “Replace in current buffer” and click the “Replace All” button. Your document should now have all text removed and look like this:

```
97  413 142 82
175 470 245 99
120 504 120 180
67  0   96  85
59  280 160 80
65  0   383 273
65  404 147 131
87  430 121 99
56  492 200 170
128 500 135 82
114 540 280
47  629 174 414
59  100 200 300
100 500 280 220
128 399 217 84
40  144 130 130
125 345 151 62
58  360 81  69
45  363 150
102 444 170 105
42  280 77
84  511 258 232
82  386 51  121
60  430
117 117 154 232
159 597 230 92
58  377 130 92
67  348 144 130
185 700 350
75  380 163 90
271 135 340 280
```

How does this regular expression work? How would you modify it to remove the numbers but leave the text?

Paste this data into your Excel sheet to the right of the Wikipedia data. Give each column a meaningful header name (e.g. “Number of Fellows”, “Number of Undergrads”, “Number of PhD Students”, and “Other Grad Students” in that order).

5. Compute Wine Per Person

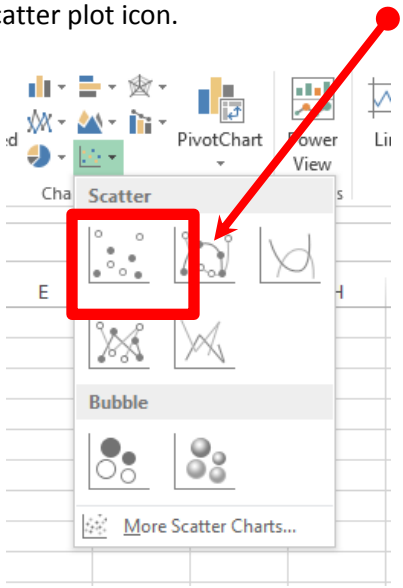
Add a column to the table giving the total number of individuals at the college (add the columns from the Cambridge University site). Notice that the Wikipedia page and the Cambridge administration page count the students differently. **Use the data from the Cambridge administration page**.

Add another column to figure out the cost of the wine bought at each college per person at the college. Remove the value for Lucy Cavendish College, which did not provide the information.
Which college spent the most? How much was it? Which spent the least? What is the average amount spent on wine per person across the whole college system?

6. Add Plots to Explore The Relationships Between Columns

Add plots to compare the amount spent on wine per person with the founding date of the college. Add another plot to compare the amount spent per person with the value of the colleges' net assets.

To add a graph, go to the "INSERT" tab at the top of Excel. Highlight the two columns of interest by clicking and dragging across the cells of the first column, and then doing the same for the second column while holding down "Control". Once the two columns are highlighted, add the graph by clicking on the scatter plot icon.



Double click on "Chart Title" to give a meaningful title to each plot.

Your scatterplots should look something like this:

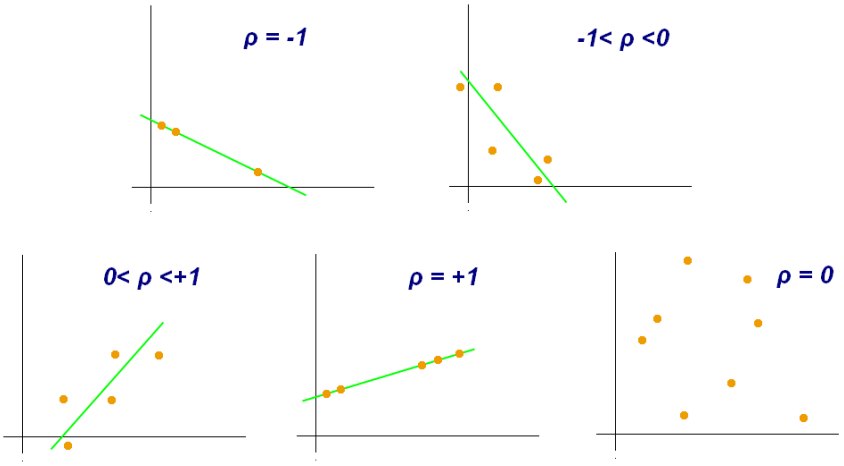
If your results are different, that is OK as this data is different from the one you would have used. Just ensure that the cells with no data are truly empty (Not a 0, "N/A" or a string with whitespace). Also ensure that you are selecting the correct columns and in the correct order.

7. Compute the Numerical Correlations

Add a row at the bottom of the table to give the correlation between each column and the amount spent on wine per person. Each cell should contain a formula similar to:

=CORREL($Q2:$Q32,B2:B32)

Here, "$Q2:$Q32" is an absolute reference to the expenditure on wine per person data and "B2:B32" is a relative reference to the data in the column above this correlation computation. Note that the column letters are dependent on where you created your "Wine Per Person" column.
Correlation coefficients can range from -1 (anticorrelated) to 1 (completely correlated). A coefficient of 0 means there is no linear correlation. See [5] for more details on the Pearson correlation coefficient.



You may find correlation coefficients **close** to the following but again, your data is different so do not be too concerned if it is not similar:

- **Wine spending per person vs founding date of college:** -.64 (older colleges tend to spend more)
- **Wine spending per person vs value of net assets:** .44 (richer colleges tend to spend more, but this is a weaker correlation).

Look at the other correlation coefficients, what else do you notice?

8. Submit Your Work

Save your Excel document as "Lab2" and submit it via OWL

[1] http://www.telegraph.co.uk/education/universityeducation/10587020/Cambridge-University-spends-3m-on-wine.html (Accessed: 19 Jan 2021)

[2] http://www.telegraph.co.uk/education/educationnews/10587171/How-much-Cambridge-colleges-spend-on-wine.html (Accessed: 19 Jan 2021)

[3] http://en.wikipedia.org/wiki/Colleges_of_the_University_of_Cambridge (Accessed: 19 Jan 2021)

[4] http://www.graduate.study.cam.ac.uk/colleges/key-facts-and-figures
(Accessed: 19 Jan 2021)

[5] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient (Accessed: 19 Jan 2021)