**Turning Unstructured Data into Structured Data**

**Overview**

Today’s lab is all about transporting unstructured data from one system to another and from one format to another. Other common terms for this process include: data ingestion & conversion, or data extraction and transformation.

**Exercise 1: Energuide Reports**

A property company had energy efficiency testing done for five houses. Each report is two pages and contains a wealth of data. Now they need to use this data to set priorities for energy efficiency upgrades. The problem is that the data is trapped in paper and PDF reports. You have been hired as a data entry clerk to transport this data from paper/PDF into an Excel spreadsheet.

|  |
| --- |
|  |

The property company wants you to extract 12 pieces of data from each of these reports into a single spreadsheet.

|  |  |  |  |
| --- | --- | --- | --- |
| 1. | House number | 7. | Third recommended upgrade |
| 2. | Overall consumption in GJ/year | 8. | Fourth recommended upgrade |
| 3. | Natural gas consumption in GJ/year | 9. | First upgrade GJ/year saved |
| 4. | Electricity consumption in GJ/year | 10. | Second upgrade GJ/year saved |
| 5. | First recommended upgrade | 11. | Third upgrade GJ/year saved |
| 6. | Second recommended upgrade | 12. | Fourth upgrade GJ/year saved |

**Part 1 Setting up the Excel sheet**

**Step 1: Open a new Excel workbook**

**Step 2: Set up the columns for our data**

Type the names of each of the data fields listed in the bulleted list above into the columns of your sheet. When you are done, you should have column names in the first row of columns A thru L. It should look like this:



**Step 3: Save your spreadsheet**

**Step 4: Spread out the columns so that we can read the column names**

Highlight columns A thru L. Position your mouse curser on the line to the right of the letter L and double click. It will change from this:

Table, timeline

Description automatically generated

To this

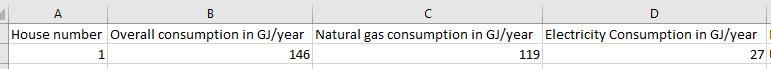


**Step 5: Don’t forget to save your spreadsheet**

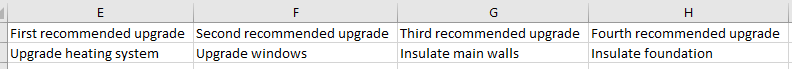
**Step 6: Open your PDF report for House 1**

Find and enter the data under the column headers you just created. This is what the twelve columns should look like when you are done

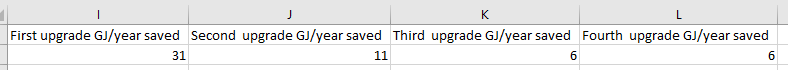
Columns A thru D



Columns E thru H



Columns I thru L



**Step 7: Save your spreadsheet but leave it open**

**Part 2 Create an Excel form**

Entering data into a spreadsheet manually is both time-consuming and prone to errors. A better approach is to create an Excel form. This approach also mimics the sort of interface you might encounter if you are asked to enter data into a database form.

**Step 8: You need to enable the “Form” button in the quick access toolbar**

You have to do this before you can create an Excel form. Click on the down arrow symbol in the quick access toolbar and select “More Commands”

Graphical user interface, application, table, Excel

Description automatically generated

**Step 9: A new menu will appear**

Change “Popular Commands” to “Commands not in the Ribbon.” Scroll down to find “Form.” Click “Add,” make sure that “Form” is now on the list on the right side and click “OK”

Graphical user interface, application

Description automatically generated

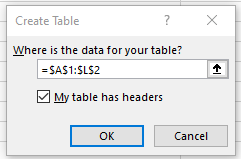
**Step 10: Clicking OK should take you back to your regular spreadsheet screen**

A new ‘Form’ button has been added to your Quick Access Toolbar

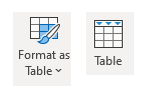


**Step 11: You need to change your spreadsheet into an Excel table, before you can make a form**

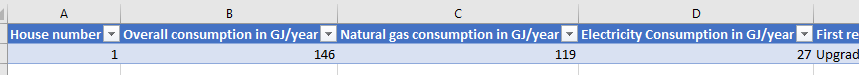
You will learn about the differences between spreadsheets and tables on another day. To make it a table click anywhere in your data and hold the Ctrl key while typing the letter T. This will bring up a tiny menu. Make sure there is a checkmark next to “My table has headers” and click “OK”



Note that there are at least two other ways of changing a spreadsheet into a form. You could select the “Format as Table” button on the “Home” ribbon or the “Table” button on the “Insert” ribbon

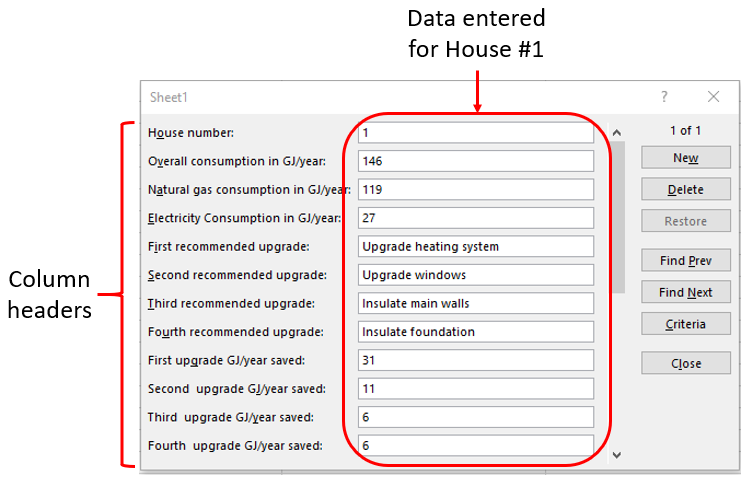


Whichever method you choose, your spreadsheet should now have down-arrows in the headers and coloured cells



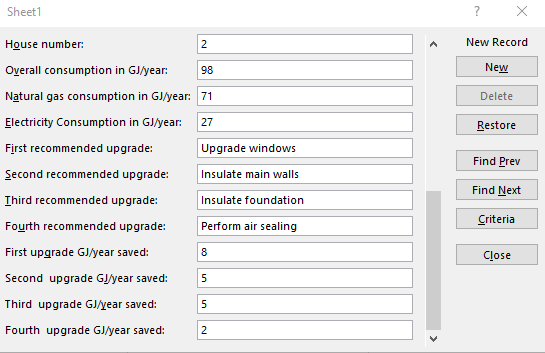
**Step 12: We are now ready to create the form**

Click on any cell containing data and then click on your new Form button in the Quick Access Toolbar. Here is what you should see:



**Step 13: You are now ready to enter the data for the remaining four houses**

Open the PDF file for House 2. Click “New” on your Excel form and enter the data for house 2. As you start entering data, note that you can use the Tab key to move down the list as you enter data



**Step 14: Repeat step 13 for Houses 3, 4 and 5**

Then click ‘Close’ in your Excel form. The Excel table will automatically update to place all the data into the appropriate cells.

**Step 15: Save and close your Excel workbook**

**Exercise 2: Words as Data**

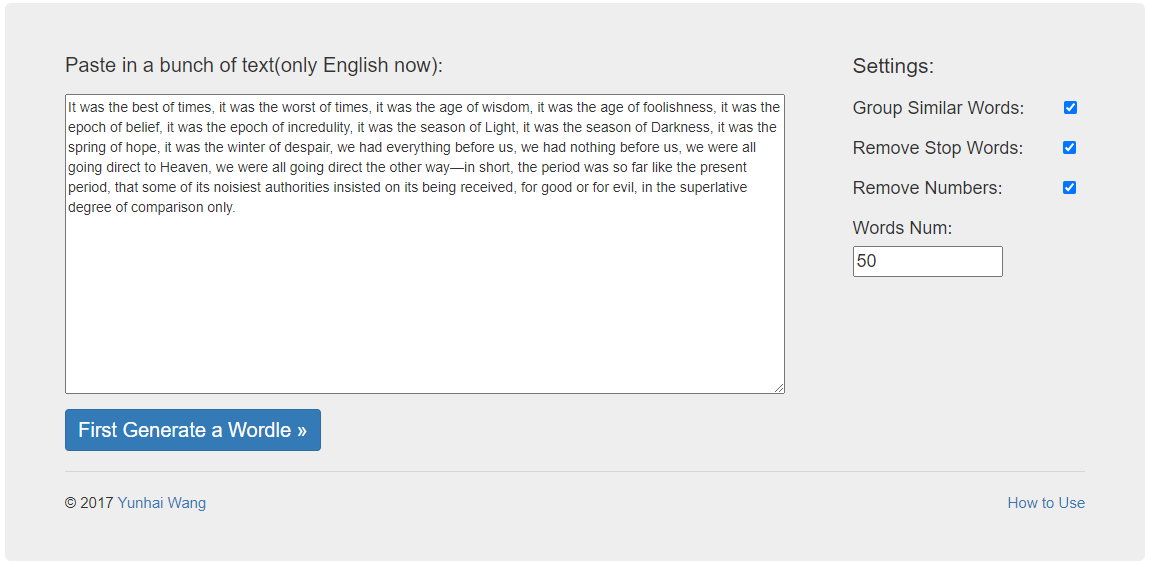
In this exercise you will transport some text on to a website for analysis.

**Step 1: Go to** [**http://www.edwordle.net/**](http://www.edwordle.net/) **and click on “Create”**

**Step 2: Copy the first paragraph from “*A Tale of Two Cities*” by Charles Dickens into the box on edwordle**

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

**Step 3: Then click on the blue button to generate your wordle**



**Step 4: Repeat steps 1 through 3 a few times**

Try changing the settings each time. What happens to your wordle when you change the settings? What do you think is going on behind the scenes?