Quick introduction to Scraperwiki and Plot.ly

This worksheet provides a brief introduction to using the scraperwiki and plot.ly to extract data from a PDF and then create a rich visualisation of the data.

Data locked inside a PDF is hard to share and for people to analyse. In this exercise we are going to unlock some data from a PDF and create a rich online visualisation from the data that others can use and build on top of.

The datasets used for this example are available openly from data.gov.my.

http://data.gov.my/view.php?view=57

Step 1 - Select your data

There are many pages and tables in the source PDF. We need to select one table to use in this exercise. You can choose any one, but this example uses Table 1.2 from page 13.

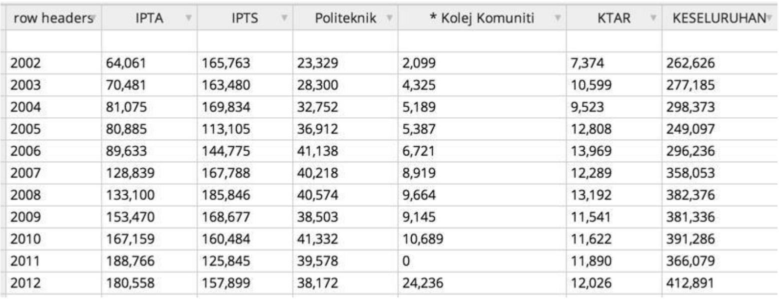
In order to use scraper wiki effectively you should separate this page into its own individual PDF. If you do not have a tool that allows you to do this then there is a copy of the page available from the course website.

Step 2 - Extract a usable table (scraperwiki)

Having prepared your source PDF, visit scraperwiki.com and click the "**get tables from PDFs**" button.

Simply upload the PDF and download the resultant spreadsheet.

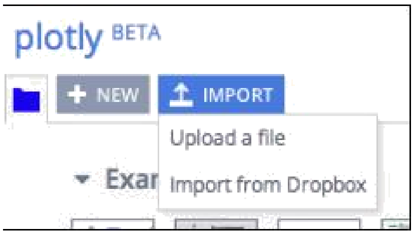
Open your table in a spreadsheet tool and remove all extraneous data such as header and footer text. Also remove the % change rows in the table as these are not required and can be re-calculated. Once done your table should look similar to the table on the right (if using the example).



row headers	IPTA	IPTS	Politeknik	* Kolej Komuniti	KTAR	KESELURUHAN
2002	64,061	165,763	23,329	2,099	7,374	262,626
2003	70,481	163,480	28,300	4,325	10,599	277,185
2004	81,075	169,834	32,752	5,189	9,523	298,373
2005	80,885	113,105	36,912	5,387	12,808	249,097
2006	89,633	144,775	41,138	6,721	13,969	296,236
2007	128,839	167,788	40,218	8,919	12,289	358,053
2008	133,100	185,846	40,574	9,664	13,192	382,376
2009	153,470	168,677	38,503	9,145	11,541	381,336
2010	167,159	160,484	41,332	10,689	11,622	391,286
2011	188,766	125,845	39,578	0	11,890	366,079
2012	180,558	157,899	38,172	24,236	12,026	412,891

Step 3 - Import into plot.ly

Register yourself at account at **http://plot.ly**, create a **new project** and **import your data**.



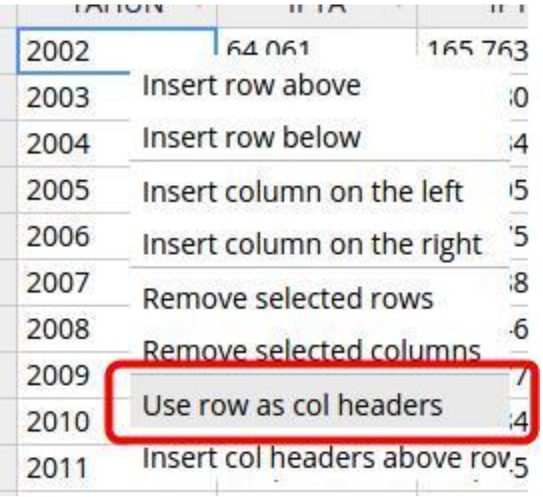
Step 4 - Tidy the data

Ensure that you have the data table displayed on your screen. In plot.ly this is named the "grid layout" and can be accessed via the grid icon.

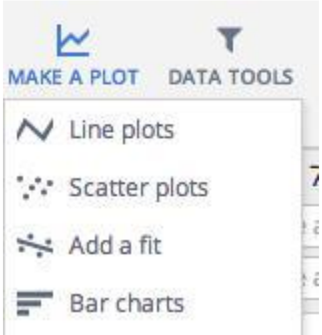
With the data displayed, it is very important to ensure that plot.ly has detected the correct column titles. For the example to work we need to switch the x and y axis so that the years are listed down the side and not along the top. This can be done by pressing the switch button.

In order to use row one as the column headings, right click any cell in row one and select "**Use row as col headers**". Once done right click row one again and this time remove it.

At the same time remove any other extraneous data in other rows.

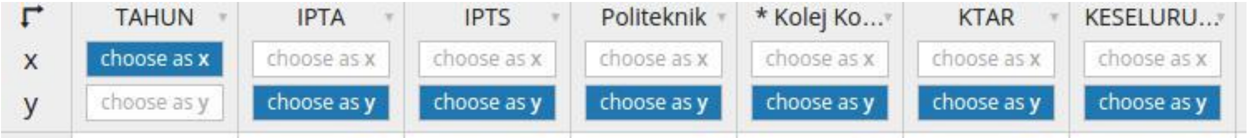


Step 5 - Make a plot



At this point you can create a plot by clicking the **make a plot** button. For the purpose of this exercise we are going to create a **Line plot** to show the change over the years. You may have to do this operation a couple of times in order to see the "choose as x" and "choose as y" selectors in the table.

Select the first column as x and the rest as y and click the "**Line plot**" button to see your interactive plot.



	TAHUN	IPTA	IPTS	Politeknik	* Kolej Ko...	KTAR	KESELURU...
x	choose as x	choose as x	choose as x	choose as x	choose as x	choose as x	choose as x
y	choose as y	choose as y	choose as y	choose as y	choose as y	choose as y	choose as y

Step 6 - Sharing you plot

Before sharing your plot, use the plot.ly toolbar to add axis titles and customise the colours used.

Once done, click the share button, give your plot a name and make it publicly available.

You could now simply send people a link to your plot, however we are going to embed this page in the website created as part of the **Publishing Data in Github** worksheet. In order to embed your visualisation you will need to have completed steps 1-3 inclusive from this worksheet.

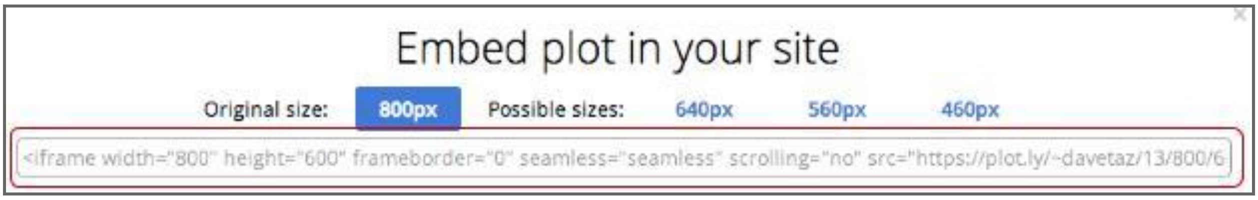
In a separate tab, log into GIthub and open your data publication project. From the list of files click on **index.html** to open it. From here click the **edit** icon (pencil) and scroll down to find the following line:

{% include data.html %}

Above this line add the following:

```
<h2>Visualisation exemplar from data</h2>
```

Keep this window/tab open and **return to plot.ly** and click the embed button in the sharing settings. In the window that appears, copy the code block and paste this below the `<h2>`element you just created in Github.



```
<iframe width="800" height="600" frameborder="0" seamless="seamless" scrolling="no" src="https://plot.ly/~davetaz/13/800/6
```

In Github, commit your change (with useful comment) and then browse to your website to view your new embedded visualisation. The link to your website is available via the **settings** menu of the repository.

Note that your visualisation is interactive, you can zoom in and select specific regions. There is also a direct link to the data used for the visualisation and the code used to generate it, making it fully open.