Data Science Club

**Data science for policy workshop #1**

Importing and describing data

What software do I need?

* Anaconda (specifically for running Python in the Jupyter environment): <https://www.anaconda.com/distribution/#download>
* GitHub. The Data Science Club portal: <https://github.com/hksdatascience/bootcamp>

What is a module

Modules are code that other people have written for a variety of functions that you can use in your own projects, giving you a variety of functionalities. The one we will be using most is **pandas**, which is used for working with tabular data.

To import it into our environment run this code:

> import pandas as pd

(The *as pd* part is the optional nickname that we give it, but as we will often call this module it is useful to shorten it)

Reading files

First we need to get our data. Pandas can handle many different formats. To save our table we can set it in a variable called *df* (short for dataframe)

> df = pd.read\_csv(‘file.csv’)

> df2 = pd.read\_excel(‘file.xls’)

> df3 = pd.read\_json(‘http…’)

Understanding our data

* Calling on the top or the bottom of our dataframe:

> df.head()

> df.tail()

* Printing a list of our columns

> df.columns

* Calling on a specific column

> df[‘variable name’]

* Getting the shape (i.e. the number of rows and columns) of our table

> df.shape()

* Getting important metadata (types, number of entries, etc.)

> df.info()

* Getting descriptive statistics

> df.describe()

* Describing non-numeric variables

> df.describe(include=[‘O’])

* Counting unique values in a column

> df[‘var1’].value\_counts(dropna=False)

* Grouping the function and applying a function

> df.groupby([‘var1’, ‘var2’]. mean()