

Reproducible Analysis

The entire *Reproducible Analysis* must be written in Python as a Jupyter Notebook. You are free to draw on concepts and methods covered in both *Quantitative Methods* and *GIS*, but must still write the code in Python (e.g. adapting something from R in the GIS module to Python). Please ensure that your submission includes: the module name, your group's student ids, and the title of your *Policy Briefing*.

How We Measure Reproducibility

We will assess reproducibility by selecting “Restart Kernel and Run All” using the `sds:2022` Docker environment. If you have made use of another Docker image then you must clearly signpost this at the start of your notebook so that we know to select a different image. We will *not* install libraries ‘by-hand’ in an *ad hoc* manner order to test the reproducibility of your work.

Reproducibility

To ensure reproducibility, markers must be able to select **Kernel > Restart Kernel and Run All Cells...** and **reproduce your *entire analysis***. This includes downloading and extracting data, cleaning, transformation, clustering... charts, tables, etc. If you need to provide supplementary or partially-processed data then you can provide this via Dropbox, OneDrive (public sharing link), or some other *robust* cloud solution that will be accessible from the marker's system.

If you have made use of one or more libraries that are not part of the Docker image then you can install these as part of your Notebook code using `! pip install`; however, if you take this approach then you should *also* ‘place nice’ by checking first to see if the library is already installed using `try... except` code that you can find on Stack Overflow and elsewhere (you will need to look this up).

Data and Resources Used

It is also up to you to ensure that all relevant data are available via a valid URL: for downloading and running. You may host your data anywhere: a GitHub repo or a

Dropbox link or some other resource but note that we may not be able to access resources placed on Chinese web servers so please bear this in mind.

i Time-Consuming Code

If your analysis has a particularly time-consuming stage (e.g. Named-Entity Recognition or Part-of-Speech tagging) then you can provide partially-processed data: comment out the code up to the point where you have generated the ‘expensive’ data set but leave it in the notebook. That way we can see how you generated the data without it being part of the **Restart Kernel and Run All Cells...** reproducibility stage.