**Assignment requirements:**

* Chosen **datasets** should be publicly accessible datasets
  + There are numerous resources on the WWW with publicly available datasets (see “Sample data sources” section
* at least two separate datasets
  + The 2 datasets you choose should be completely different to each other although they *can* belong to the same domain if you intend to combine them in your analysis. Cases like (but not limited to):
    - * One dataset being a derivative of the other,
      * Both being subsets of the same original dataset,
* Each dataset should have at least 1,000 records (rows).
* You must provide evidence in your report that you are authorized to use the dataset(s) that you have chosen.
* The word count not less than 2,000 words, and not more than 2,500 words (not counting R code).
* For each of the chosen datasets you are required to compile a report of your analysis.
* The main deliverable is a report that provides significant insights into the datasets that you have chosen to analyse.
* Your report should provide at least four unique insights based on your data analysis. Examples of insights might include relationships, trends/patterns, correlations, models based on the data, visuals, and statistical analyses
* Your project report should discuss the challenges that you encountered while handling your chosen datasets and the means and mechanisms you implemented to overcome these challenges.
* All deliverables should be compiled into a project report document – Word document, along with all programming code elements in an appendix
* R scripts and additional files – in a separate zipped file

It should contain *at least* the following resources:

• Your R code as 1 or more .r file(s),

• Your 2 datasets in the state they are input into your R code.

Please use relative filepaths in your code, such that unzipping the archive allows all code to be run without any need to edit the filepaths or move the datasets around.

Other included files can (but are not required to) be intermediate dataset outputs, graphs/plots/etc as graphics, or whatever other artefacts you deem appropriate

* Pre-processing done in R
* **Referencing and plagiarism**

You are expected to cite any code you use that is not strictly your own.

• If you’re reusing a whole script or class, you should reference all the appropriate details (URL, original author, date, etc.) in a large comment block at the top of the file.

• If you’re reusing a code snippet, you should reference its origin URL and author in inline comments just above the snippet.

**Sample data sources see below.**

