

Data Analysis Skills - Assessment 1

Objective: Create a Statistical Analysis Plan (SAP) based on data obtained by web scraping using the **rvest** R package and review the SAPs of others in your group.

Deadlines:

- For data and document upload: **17:00 Friday 28th May**
- For peer review: **17:00 Friday 4th June**

Contribution to your final grade: 5% total (3% for submitting your files and 2% for completing the peer review exercise for each member of your group)

Overview: You are required to obtain a dataset of interest to you by ‘scraping’ data from a website of your choosing (and for which you have permission) using the **rvest** R package. You will then pose at least 2 research questions that you could explore using the dataset. You will then submit your data and a document that summarises your dataset, lists your research questions and includes a Statistical Analysis Plan (SAP) which outlines how you would explore your research questions.

Your submitted data and document will be peer reviewed within your group and you will review the data and documents submitted by other members of your group.

Task Details

Task 1 - Obtain a dataset of interest to you by ‘scraping’ data from a website of your choosing using the **rvest** R package and identify two research questions that could be explored using the dataset. Before you scrape the data, ensure that you have permission by using the **paths_allowed()** function.

Once imported into R, store your data as a **tibble** with a meaningful name and give the variables appropriate names and ensure the variables are the correct data types in R (i.e. character, numeric, integer, factor, logical, etc.). Export your data to an Excel (.xls) file and upload this to Aropa (see below).

Your data must meet the following requirements:

- The data should have between 50 and 500 observations.
- The data must **not** be scraped from any of the websites used in the course videos/materials (e.g. www.imdb.com and www.opensecrets.org) **with the exception of Wikipedia which you may use.**

Task 2 - Produce a Word or PDF document containing:

- The URL (web address) of the data
- The context of the data and its variables
- The content of the data as summarised by the output of the **str()** function applied to the **tibble** containing the data
- The research questions you want to explore using the data, and
- A SAP for exploring the research questions using the data.

There is a template for the document available on Moodle and your SAP must be no longer than 2 A4 sheets in length.

Submission Instructions

The Excel file and your Word/PDF document must be submitted to **Aropa** for peer review via the link on the DAS Moodle page. Use your GUID (e.g. 1234567m) and GUID password to log into the system.

Your dataset must be saved as an Excel (.xls) document with the file name **Data.xls** and your document must be saved as a Word/PDF file with the file name **SAP.doc/SAP.pdf**. Please **DO NOT** include your name or student number in the file names or the actual document, this will help keep the peer review process as anonymous as possible.

When you are submitting your files you will be asked to select a tag. Please **select your group number** e.g. if you are in group 1, select “1” in the tag field. The group allocations can be found on the [DAS Moodle page](#).

BOTH submissions must be made by **17:00 Friday 28th May**.

Peer Review Instructions

- **After the deadline** for the AROPA upload you will be able to peer review the submissions of everyone in your group using AROPA.
- Log into AROPA via the same link you used to uploaded your files.
- You will then see each of your group members' files that you will give constructive feedback on by answering questions in AROPA.
- Once you have completed reviewing your group members' submissions you will be able to see their reviews of your submissions.

All peer reviews must be completed by **17:00 Friday 4th June**.

Any technical difficulties **MUST** be reported to Craig (craig.alexander2@glasgow.ac.uk) before the submission/reviewing deadlines.