Applied Health Data Science Summative Assessment 2:

Data wrangling and visualisation task

(50% of Unit Marks)

## Key deadlines for your diary (non-negotiable):

Assessment Submission: 12 noon on Wednesday 17th January 2024

## 1. Assessment Instructions

The second summative assessment for this unit involves following reproducible coding principles in developing a code pipeline to access the data set provided, explore it and present interesting findings visually.

Your submission should be a **zipped, executable, reproducible research compendium** (<https://the-turing-way.netlify.app/reproducible-research/compendia.html>) in the same form as the one you produced during the formative assessment, version controlled using Git, and including:

* A **readme** text file introducing the compendium (between 100 and 300 words).
* The **raw** data (in a directory called **raw**).
* The **clean** data (in a directory called **clean**). The visualisations should be generated using this copy of the data.
* The **code** (with comments) used to conduct all steps to generate the visualisations from the data provided. This should include any further readme files to help document your code, and be provided in a directory called **code.**
* An interactive Shiny **visualisation** produced by the code, in a directory called **visualisation.**
* A record of the computational **environment**.
* A **pipeline** in bash or snakemake to run the code and produce the report
* A **report** (in Word or PDF format) of no more than 700 words, with the following sections:
  1. **Research Question:** the non-trivial question you are intending to address with this report (a trivial question would be asking whether weight correlates with height, for example, because you can tell that will almost certainly be true without looking at the data).
  2. **Data Management:** a description of the data and the steps you undertook to pre-process it (e.g. data cleaning).
  3. **Data Visualisation Approach:** describing the methods you used to visualise the data, the theoretical reasons you chose these approaches, and instructions for viewing the Shiny visualisation.
  4. **Results:** two ggplot2 visualisations embedded in text describing your interpretation of them (i.e. what does each plot show us that we did not know before). This section should also include interpretation of the interactive Shiny visualisation that can be viewed separately.
* The **.git** hidden folder (with code tracked from the beginning of the project).

Make sure the first thing you do is to set up the version control with git in the project’s directory, so that all stages of the development of the project are tracked. All code should be set up to run from the bash command line (either locally or on BlueCrystal). Your code should be easy for others (including the markers of this assignment) to run.

**The zip file containing the compendium should be named AHDS\_assessment2\_[your student number].zip.**

Where possible, the compendium should not contain identifiable information, such as your name, or University or GitHub usernames (apart from in the .git folder).

**You will be working individually on this assessment. Any evidence of copying between students will be penalised in your final mark.** We have set up a Padlet on blackboard for you to ask questions about this task. This is so that everyone has access to the answers. All questions regarding the task should either be asked during teaching sessions or via the Padlet.

Assessment dataset

You are provided with a file called nhanes\_ahds\_23\_24.zip. It contains three data files containing data from the NHANES survey (<https://www.cdc.gov/nchs/nhanes/about_nhanes.htm>). Documentation on the specific datasets we have provided can be found here – <https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination&CycleBeginYear=2005> (BMX\_D)

<https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&Cycle=2005-2006> (DEMO\_D)

<https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Dietary&Cycle=2005-2006> (the SEQN and BMXBMI fields from FFQRAW\_D)

You do not need to work with all these data, but should choose a subset that you can use to answer an interesting question in your report.

## 2. Intended learning outcomes

This assessment covers unit intended learning outcomes 3, 4 and 5:

(3) Write an analysis pipeline using Linux and R

(4) Produce R and web-based data visualisations

(5) Identify and implement approaches for conducting reproducible research

## 3. Good academic Practice

This is a piece of COURSEWORK that contributes to your Unit mark and you can:

* + Use resources to support you in completing your answer.
  + Draw upon a range of accepted resources including, your own notes, lecture slides/recordings, course material, textbooks, journal articles, online resources. ALL work should be written in your own words.
  + Ask for help from your personal tutors or academic lecturers if you do not understand an aspect of the coursework.
  + Please post any questions on the PADLET available for this assignment on Blackboard.
  + Broad discussion with your tutors, fellow students, friends and family on the assessment topic and your ideas/approach may help you to further your knowledge and understanding.
  + Use your network of family and friends to gain support and encouragement during the assessment period.

Please remember this is a formal assessment and you should behave in a manner consistent with our values. This means you cannot:

* + Allow others to directly contribute to your written answer by revising or adding to the academic content. This is collusion and is against University Regulations.
  + Share your assessment with others or ask others to share their work with you.
  + Copy and paste any material (text, images, coding, calculations) from other sources, including teaching material and shared revision notes directly into your answer without appropriate acknowledgement. This is plagiarism (see section 5)
  + Pay another person or company to complete the assessment for you. This is contract cheating and is against University Regulations.

Note that this means that in this specific case the reproducible research practices you employ should not include approaches such as pair programming or code review by a second person.

## 4. Assessment submission

* The assessment should be submitted via Blackboard by **12 noon on Wednesday 17th January** **2024**. There are two submission points, one for the complete zip file (including the report), and a second Turnitin submission point for a copy of the report by itself (so that it can run through Turnitin).
* Please label your zip file submission as follows: AHDS\_assessment2\_[your student number] e.g. AHDS\_assessment2\_01234.zip.
* Please label the copy of your report as follows: AHDS\_assessment2\_report\_[your student number] e.g. AHDS\_assessment2\_report\_01234.docx
* Please include an assessment cover sheet with your assessment – this is available on Blackboard.

**Failure to submit on time will incur penalties unless you have an approved extension.** Please see the Programme Handbook for further information on penalties. Please see the Programme Handbook on Blackboard for details on the extension and extenuating circumstances procedure, and for the penalties for late submission of coursework. If you encounter problems with your submission, or if you are likely to have a problem submitting on time, please email the course administrator as soon as possible [brms-msdscourseadmin@bristol.ac.uk](mailto:brms-msdscourseadmin@bristol.ac.uk).

## 5. Marking

Your research compendium will be marked according to three categories, each with a portion of the marks allocated to it. Within each category, marks will be assigned based on the marking scale in the programme handbook.

* The report (50%)
* The interactive Shiny data visualisation (20%)
* Application of reproducible research practices, including the pipeline, version control, organisation and readme documentation (30%)

## 6. Referencing, Copyright and intellectual property

It is important that the coursework you submit is your own work. All written assessments will be checked by Turnitin for issues related to plagiarism, collusion, and cheating. Please see the Programme Handbook for important information on academic integrity. You are also able to re-listen to the course Welcome Week academic integrity talk at any time or view the guidance on the University’s website [www.bristol.ac.uk/students/support/academic-advice/academic-integrity/](http://www.bristol.ac.uk/students/support/academic-advice/academic-integrity/)

Copyright and intellectual property rights are also important issues to be aware of when using the work of others in your coursework. This is not just about ensuring that you correctly reference everything, but you also need to be sure that you have *permission* to re-use this work. Examples of this might be displaying a figure you have taken from someone else’s work or using an existing questionnaire. If you have any concerns about copyright issues, please speak to the unit lead in advance of submitting your assessment.

You may include photographs or scans of your own hand-drawn, labelled diagrams or calculations. We would advise you to generate your own diagrams but if you include diagrams or pictures that you have not produced yourself, or are modified versions of existing images, you should ensure you reference them appropriately.

## 7. Wordcount

The word limit for the report is 700 words. This includes:

* + **All** text including in-text titles and headings
* Figure legends
* Text in tables

All in-text citations

The word limit does not include:

* + Cover sheet
  + Text in data visualisations (both static and interactive)
  + List of references

You should include two static figures and an interactive visualisation.

All tables (if any) and static visualisations must be included embedded in the relevant place in the text of the report. You should include instructions on how the interactive visualisation can be viewed.

You must provide the word count of the report section of the submitted work on the cover sheet accompanying your submission.

Exceeding the word limit will incur the following penalties. You will be informed of any penalties applied to your assessment:

|  |  |
| --- | --- |
| Coursework that exceeds the stated word limit by: | Penalty (absolute): |
| Up to 5% | 5% of total mark available is deducted\* |
| Between 6-10% | 10% of total mark available is deducted\* |
| Between 11% and 20% | 20% of total mark available is deducted\* |
| Between 21% and 50% | 50% of total mark available is deducted\* |
| By over 50% | A mark of 0 is awarded |

\*Note: the minimum mark is 0, negative marks will not be given.

Please see the Programme Handbook for further information on word count limits and penalties.