Applied Health Data Science Summative Assessment:

Health Data Science Mini Project

(100% of Unit Marks)

Key deadlines for your diary (non-negotiable):

Assessment Submission: 12 noon on Monday 2nd December 2024

1. Assessment Instructions

The summative assessment for this unit involves following reproducible coding principles to develop a code pipeline to complete a data science mini project.

Mini project description

The aim of this mini project is to explore trends in COVID-19 publications. To achieve this, you will download, process and visualize data on the text of research articles.

All tasks should be designed to run on BlueCrystal as part of a SnakeMake pipeline. Your SnakeMake pipeline will include four steps. The whole project should be tracked using Git from the start of development, and you should use Conda for package management. All code for each task should be included in your submission, along with an exported Conda environment file. The Git log should be a complete record of your work in developing the pipeline.

The first thing you will need to do is to set up version control using Git, and a Conda environment with the required packages.

Next, design the pipeline. The steps you should include are:

Step 1: Download a set of research articles on the topic of long COVID using bash.

PubMed (https://pubmed.ncbi.nlm.nih.gov/) is a free online database of research articles, including their titles, abstracts and meta data such as publication date and author list. Each article has a PubMed ID that is unique for that article. PubMed has an application programming interface (API) that allows us to access the database programmatically (by writing commands and scripts) rather than through the website.

You should retrieve the IDs of the relevant articles from this URL:
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=%22long%20covid%22&retmax=10000

You can then retrieve the article meta data for each article using a URL such as:
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=39240571

To retrieve the article IDs and then the article summaries programmatically, you can use curl commands like this (each curl command on a single line):

```
curl
"https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db
=pubmed&term=%22long%20covid%22&retmax=10000" > data/pmids.xml

curl
"https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=
pubmed&id=${pmid}" > data/article-data-$pmid.xml
```

You should make a script to retrieve these data programmatically, and integrate these scripts into your pipeline. When retrieving the article data, you should include a pause between downloading each article file so that the PubMed website isn't overwhelmed with requests. You can do this using the sleep command, e.g.:

```
sleep 1 # this will pause for 1 second
```

See information about the E-utilities (the API through which PubMed can be accessed programmatically) here – https://www.ncbi.nlm.nih.gov/books/NBK25501/

Step 2: Process the downloaded XML files, programmatically using bash or Tidyverse:

- Extract the year and title of each article and store these in a tab separated file with three columns: PMID, year, and title.
- Remove any xml tags (e.g. and) from the titles.
- Remove any rows for articles that do not have a title
- You may also choose to extract and use the article abstracts or MESH terms, in addition to, or instead of the abstract title.

Step 3: Process the titles using the tidytext package:

Your processing might include:

- Removing stop words and digits
- Reduce words to their stem

Step 4: Produce an informative data visualisation to show the change in the text of the article titles (or other fields you have extracted) over time:

For example, this could be changes in frequency of particular words over time, changes in the frequency of MESH terms extracted from the article XML, or a more advanced analysis you have encountered in your wider reading, such as changes in article topics over time using Latent Dirichlet Allocation (LDA) topic modelling.

Submission points

There are two submission points for this assessment, one for the code and one for your report.

1. Code

Your code should be submitted as a **zip file** and include:

- An empty directory called **raw** where the raw data is stored (you should not include the data, just the directory where it should be placed).
- An empty directory called **clean** where the clean data generated by your scripts will go.
- The **code** used to conduct all four steps of the analysis (with readme files and comments). The pipeline should be designed to run on BlueCrystal using Bash scripts, R code, SnakeMake and Conda.
- The **.git** hidden folder (with code tracked from the beginning of the project).

Your code should be designed so that a marker can upload it to BlueCrystal and it will run.

The zip file containing your code should be named AHDS_assessment_code_[your student number].zip and be submitted to the "code" submission point.

2. Report

The report should be completed using the **template** provided.

Report sections:

1. **Data processing**: A description of the steps you undertook to pre-process the data (e.g. reformatting, data cleaning).
2. **Data visualisation with description and justification of approach:** One static visualisation produced in R and embedded in text describing the question you intended your plot to explore, the methods you used to visualise the data, the theoretical reasons you chose this approach, and your interpretation of what the plot shows.
3. **Reflection on ethical and governance considerations:** Imagine that you were going to request a data set from the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort study (https://www.bristol.ac.uk/alspac/) to explore how participants' health changed over the course of the pandemic. This data set will contain individual-level sensitive medical data from people living in Bristol. What ethical or governance considerations should you take into account?
4. **Reflection on data management:** Now imagine that ALSPAC has approved your project to work with the data set you requested. How do you plan to store and manage the data in an appropriate way?

The word document (.docx) containing your report should be called AHDS_assessment_report_[your student number].docx and be submitted to the "report" submission point.

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.

2. Intended learning outcomes

This assessment covers the following unit intended learning outcomes:

1. Formulate a data management plan for a research project.
2. Critically evaluate the ethical and governance considerations that are important for health data scientists.
3. Implement an analysis pipeline using Linux and R.
4. Develop theory-informed visualisations to explore and explain patterns in health data.
5. Identify and implement approaches to ensure your research is reproducible.

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

Assessment should be submitted via Blackboard by **12 noon on Monday 2nd December 2024**. Please label your submission as follows: AHDS_assessment_report_[your student number].docx and AHDS_assessment_code__[your student number].zip. Please include an

assessment cover sheet with each submission, as the first page of the report and in the top level of the code zip directory. The cover sheet 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.

5. Marking

Your mini project 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 (40%)
  - Sections 1 to 4 of the report will be weighted equally
- The code (60%), including:
  - Working, well organised and well documented Bash scripts and R code that completes each task in the mini project (40%)
  - Appropriate use of version control with Git and environment management with Conda (10%)
  - Appropriate use of SnakeMake (10%)

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/

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 **1,200** words. This includes:

- **All** text including in-text titles and headings, and including template text
- Figure captions
- Text in tables

- All in-text citations

The word limit does not include:

- Cover sheet
- Text in the data visualisation image
- List of references

You must provide the word count of your report on the cover sheet accompanying the report part of 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.