MA30280 Applied Data Science - CW 1

Deadline: Monday, 11th March

Contents

General Instructions

Set: 20 February 2024.

Due: 11 March 2024, 12pm.

Estimated time required: 30 hours.

Submission: On Moodle. The file to submit must be a compiling RMarkdown (.Rmd) file and an eventual zip folder containing external resources used in the analysis (for example, Excel files, images, etc.). The Rmd file should be named "`MA30280_CW1_CandidateNumber.Rmd`", and it must **NOT** contain the author information in the YML header to guarantee anonymity of the candidate. The RMarkdown output (i.e., pdf, html, word) should include **no** computer commands and **no** raw output. The `R` code chunks in the Rmd file should be commented. This will not be marked for credit, but it might be checked to see if it is unclear what you have done in your report, so ensure that it is intelligible and does not contain redundant material.

Conditions: The work submitted is individual and personal. Your lecturer may answer generic questions about methodologies and computing relevant to the coursework, but not specific questions about specific analyses. To keep things fair, questions relevant to the whole class will not be answered individually, but will be answered on the Moodle forum. Do not ask other members of staff, post-graduates or students for help.

Value: 50% of the unit mark. See the **Marking Scheme** section for more details.

Length: No more than 5 pages. Appendix and references, if any, are **not** included in the page-count.

Support and advice: For any doubt on the coursework assignment, contact me during the Q&A sessions in week 5, or via email at ib641bath.ac.uk. For IMCs or extensions, contact your Director of Studies.

Feedback: You will receive feedback within a maximum of three semester weeks following the submission deadline. The feedback will consist of your marked work and an overall feedback document commenting on the assessment across the cohort.

Late submission of coursework: If there are valid circumstances preventing you from meeting the deadline, your Director of Studies may grant you an extension to the specified submission date, if it is requested before the deadline. Forms to request an extension are available on SAMIS.

- If you submit a piece of work after the submission date, and no extension has been granted, the maximum mark possible will be the pass mark.
- If you submit work more than five working days after the submission date, you will normally receive a mark of 0 (zero), unless you have been granted an extension.

Academic integrity statement: Academic misconduct is defined by the University as "the use of unfair means in any examination or assessment procedure". This includes (but is not limited to) cheating,

collusion, plagiarism, fabrication, or falsification. The University's Quality Assurance Code of Practice, QA53 Examination and Assessment Offences, sets out the consequences of committing an offence and the penalties that might be applied.

Generative AI: *Type B*

Generative AI is permitted as an assistive tool for specific defined processes within the assessment and its use is not mandatory in order to complete the assessment.

In particular, under the University's Academic Integrity Statement, you "must not present content created by generative AI tools as though it were your own". Any text or code produced by generative AI must be checked for correctness and cited. In addition, you must include a short statement (max 250 words) at the end of your submission indicating: what tools you used and how you used them **OR** that you have not used generative AI tools. You should be prepared to explain anything in your submission to an examiner if asked to do so.

Your generative AI statement is **not** included in the 5-page limit.

See GenAI Assessment Categorisation.

Analysis

You are working for a humanitarian consulting agency here at Bath, and your weekly project is on *provincial data for Argentina.* This is part of a year-long program with the aim of planning public policy to help allocate resources to develop infrastructure, education, and welfare schemes. Depending on the ten economic and social indicators collected for each province, you need to partition the provinces into groups with similar development levels to optimise the delivery and the planning of possible public policies.

Variable descriptors and the data are in the related data files, "`argentina.xlsx`".

Structure your essay as follows:

1. Run some exploratory data analysis on the Argentina data. In particular,
    - Control for any miscoded variables or any missing data.
    - Compute the necessary summary statistics for all variables in your data. Create additional variables as GDP per capita and population in percent to better compare the different provinces. Display interesting bar-plots where possible differences between provinces can be detected.
    - Are the variables correlated? A correlogram might help in visualise the correlation structure between the variables.
    - Can you highlight the provinces that might need highest priority?
2. As economic and social indicators are highly correlated, first run principal component analysis (PCA) to reduce redundancies and highlight patterns that are not apparent in the raw data. After visualizing the patterns, use an appropriate clustering algorithm to partition the provinces into groups with similar development levels.
3. Write a conclusion section of your results, summarising in few sentences your findings. Discuss possible limitations of the data or the methods used, and hypothesise what further analyses can be conducted from your outcomes (e.g. supervised learning methods). In such section, spend few words to suggest also for possible policies that the Argentina government might actualise to improve the life quality in those provinces with highest priority. For example, a possible suggestion could be to propose more public funds to strengthen a particular aspect of the education system.

Marking scheme

The project will be marked out of 20 marks. There is no single correct analysis for this type of project, so you will not be marked on the basis of how close you get to some particular answer. The marks are not subdivided, but will be allocated on a combination of data science approach and justification, interpretation of results in context and presentation.

14-20 (First)

A report that could be presented with little or no revision. Analysis is sound so that conclusions are scientifically well-supported. Interpretation is reasonably mature. The project should demonstrate a clear overview of the work, without getting lost in details, and be free of all but minor statistical errors.

12-14 (2.1)

A report that could be presented after a round of revision, but without having to re-do much of the actual analysis. Some flaws in the analysis or presentation (or minor flaws in both), but basically sound. A good grasp of the data science techniques and context, so that interpretation is reasonable.

10-12 (2.2)

Major re-working required before the report could be presented, but containing some sound data science thinking demonstrating understanding of the modelling approach and its application. Reasonable presentation and organisation.

8-10 (Third)

Major flaws in analysis and presentation, but demonstrating some understanding of data science methodologies, and a reasonable attempt to present the results.

0-8 (Fail)

Flawed analysis demonstrating little or no understanding of data science, and/or incomprehensible or overly bad organisation/presentation.