MA22019 INTRODUCTION TO DATA SCIENCE - COURSEWORK 2

Contents

General instructions

Set: 9:00 on 25 April 2025

Due: 12:00 (noon) on 6 May 2025

Estimated time required: 15 hours for students familiar with the content covered in MA22019.

Submission: You should submit your answers in two different files: one PDF that contains your answers, including any R code and output, and the R Markdown file you used to produce the PDF.

Conditions: This is an individual assignment, and you should not discuss it with anyone other than the lecturer or your tutor. It should be completed during your computer lab on 29 April and in your own time.
You may use whichever R packages you find useful for this coursework, but they need to be referenced at the end of the PDF, unless the package was considered in the lectures. You can access the necessary information using the citation() function in R, e.g., `citation("dplyr")` states how you should cite the dplyr package.

Value: This assessment caries 60% of your mark for MA22019 Introduction to Data Science.

Marking: There are 50 marks available, and the marks are split across three questions. For each question you will have to perform and present a data analysis which will be marked according to the following three criteria:

- **Correctness:** Is the analysis correct? Is the choice of plots appropriate and are all R outputs interpreted correctly?
- **Presentation:** Is the overall analysis presented in a clear and concise manner? Is the analysis focused on addressing the question? Is it easy for others familiar with the content of MA22019 to follow the thought process?
  Note: The amount of white space is not used as a criteria, but the size of figures is.
- **Quality:** How well does the presented analysis address the research question? Have all assumptions been clearly stated and, where appropriate, been discussed in enough detail? Are the conclusions useful and discussed in context? Does the report appropriately highlight potential limitations.

Length: There is no minimum or maximum length for this assignment; when marking, emphasis will be placed upon clear arguments and conciseness.

Support and advice: You will be able to ask questions during during your tutorial on 29 April. There will also be an office hour on 10:00-12:30 on 30 April. Questions can also be posted on the MA22019 Padlet board. You can ask generic statistical, coding or presentation questions relevant to the coursework or the course in general, but not specific questions about how to do the coursework analyses.

Feedback: General feedback for the overall cohort will be provided within three weeks following the submission deadline. Individual feedback and provisional marks will be released on Results Day in July.

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. **It has been agreed with the Maths Director of Studies team that extensions beyond 12:00 (noon) on Saturday 10 May will only be granted in exceptional circumstances.** 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, QA53Examination 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 within the assessment, but its use is not mandatory in order to complete the assessment. Please provide a statement at the beginning of the coursework which states if you have used generative AI tools and, if so, describe for which purposes you used them. See GenAI Assessment Categorisation for more details.

Contact Details:
Christian Rohrbeck
Office: 6 West 1.18
E-mail: cr777@bath.ac.uk

Background and Questions

Background: The country of Utopia has collected large amounts of data over the past years, but it is short of data scientists who can help with analyzing that data. You have thus been approached to help them with addressing some urgent questions. You have been provided with several data sets that may be useful. A detailed data description for the different files is provided at the end of this document. There is no expectation for you to use any data other than that provided.

Reproducibility of results is crucial for the people in Utopia. Therefore, you have been provided with a R Markdown template which you have to use. You are asked to create a PDF file (using R Markdown) that contains all your answers and necessary details of your analysis, including the R code and outputs your answers are based on.

Questions: Use the data science techniques covered in the course to address the following research questions / tasks:

(1) Generative AI tools have gained a reputation for their ability to produce texts in the style of famous authors. You are asked to put this claim to the test using the work by Charles Dickens.

The file “Dickens_Fakes.csv” contains five pieces of text written by ChatGPT in the style of Charles Dickens, and “Dickens_Originals.csv” provides ten randomly selected chapters from the books *A Tale of Two Cities* and *Great Expectations.* Explore the similarities and differences between the original and artificially produced pieces of text. Your analysis should consider both the words used and the emotional intent of the texts.

What do you conclude regarding ChatGPT’s ability to produce work in the style of Charles Dickens? How confident should we be about your conclusion? **[15 marks]**

(2) The region of *Tesremos* is one of Utopia’s key sources for freshwater. Unfortunately, Tesremos has seen lower than expected rainfall over the past years, which in the longer term may threaten the country’s capacity to deliver clean water to its citizens. Due to its importance, the government of Utopia has been monitoring ground water levels at several sites across Tesremos for decades.

The government of Utopia has now approached you to analyse some of last year’s data to get a better understanding of the current situation. Specifically, they provided you with daily observations from their 102 groundwater monitoring sites across Tesremos for March and August 2024, which correspond to the beginning and end of Utopia’s dry season respectively. Instead of the actual values, the government of Utopia has decided to provide you with the differences (in percent) to historical averages. As such, a value of x for a day and site in 2024 corresponds to ground water levels being $x\%$ lower (for x negative), or $x\%$ higher (for x positive), than the groundwater levels which were observed historically for that day and site.

You are also provided with a shapefile for Tesremos, and the locations of the 102 monitoring sites. To hide Tesremos’s location, constants have been added to the latitude and longitude coordinates, but the shapes they define are correct.

There are two tasks the government of Utopia asks you to complete using the provided data:

i Explore the spatial pattern/structure and dependence of the monthly average deficit in groundwater levels for the 102 monitoring sites separately for March and August. Discuss all the assumptions you make for the analysis.

ii Produce maps which provide estimates for the monthly average deficit in groundwater levels across the region for March and August 2024. Discuss the reliability of your estimates.

[20 marks]

(3) The *Isles of Sofara* are one of the most beautiful spots across Utopia and very popular with tourists. There are a total of five islands: *Calmorra*, *Ecliptria*, *Justitia*, *Paxora* and *Solvenya*. You have been provided with a shapefile of the islands; to hide their location, constants have been added to the latitude and longitude coordinates, but the shapes they define are correct.

Over recent decades, the islands have been frequently hit by devastating tropical storms and the government of Utopia was repeatedly blamed for their poor disaster response. To address this issue, the government of Utopia set up a helpline in 2020 - every person in distress due to a tropical storm should send a text message to the number 971 with the hashtag #IoSTSHelp and, if possible, provide their location information. Since 2020, a few tropical storms hit the islands and, while the introduction of the helpline and hashtag was generally successful, the current setup requires a government employee to manually evaluate the text messages received, which leads to delays in the disaster response.

The government of Utopia has now approached you for help with the analysis of the text messages sent to 971. Specifically, for the messages with the hashtag #IoSTSHelp, they ask you to:

i Visualize the number of messages per island and identify, as precisely as possible, the areas from which the most distress messages were sent.

ii Develop an approach to identify the kind of help required. The four key categories are (a) lack of food, (b) lack of fuel, (c) lack of shelter and (d) need for medical assistance. Use your approach to provide estimates for the number of people who required help related to the categories (a)-(d).

The government of Utopia is planning to use your analysis approach for any future tropical storms. Specifically, they want to use your approach to identify the areas that need help and the kind of help required, without a government employee having to manually evaluate the text messages. Discuss the strengths and limitations of your analysis approach regarding this aspect. **[15 marks]**

Data Descriptions

You have been given access to various data sets and the UTF-8 encoding standard is used throughout:

AFINN Sentiment Lexicon.csv: AFINN sentiment lexicon:
word: Word
value: Assigned sentiment value

Bing Sentiment Lexicon.csv: Bing sentiment lexicon:
word: Word
value: Assigned sentiment ('positive' or 'negative')

Dickens_Fakes.csv: Five short stories by ChatGPT in the style of Charles Dickens:
Title: Title of the short story
Text: Line from the short story

Dickens_Originals.csv: Ten randomly selected chapters from the books *A Tale of Two Cities* and *Great Expectations*:
Title: Title of the book the chapter is from
Text: Line of text as printed in the book
Chapter: Number of the chapter

Isles_of_Sofara_Messages.csv: Messages sent to the 971 helpline between January 2020 and March 2025:
Message: Text which was sent
Longitude: Longitude coordinate of the person who sent the text
Latitude: Latitude coordinate of the person who sent the text
Island: Name of the island the message comes from

Isles_of_Sofara_Shapefile: Folder containing the shapefile for the Isles of Sofara.

Tesremos_Locations.csv: Locations of the 102 sites across Tesremos at which groundwater levels are measured:
ID: A unique identifier for each site
Longitude: Longitude coordinate of the site
Latitude: Latitude coordinate of the site

Tesremos_Water.csv: Daily deficit in terms of historical ground water levels for the 102 sites for March and August 2024:
Date: Date of the observations
Site_1,...,Site_102: Observations for the different sites

Tesremos_Shapefile: Folder containing the shapefile for the Tesremos region of Utopia.