MA22019 INTRODUCTION TO DATA SCIENCE - COURSEWORK 1

Contents

General instructions

Set: 15:00 on 21 February 2025

Due: 17:00 on 28 February 2025

Estimated time required: 10 hours for students familiar with the content of Chapters 1 and 2 in the MA22019 lecture notes.

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 25 February 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 40% of your mark for MA22019 Introduction to Data Science.

Marking: There are 45 marks available, and the marks are split across three questions worth 15 marks each. 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 analysis address the research question? Are the conclusions useful and discussed in context?

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

Support and advice: You will be able to ask questions during the lecture on Wednesday 26 February and during your tutorial on 25 February. Questions can also be posted on Moodle or 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: Feedback for the overall cohort will be provided within two weeks following the submission deadline. Individual feedback and provisional marks will be released within three weeks following the submission deadline.

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 17:00 on Tuesday 4 March 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 Chapters 1–2 of the lecture notes to address the following research questions / tasks:

(1) Several areas of Utopia are historically suffering from poor air quality. The society "Clean Air for Utopia" has measured PM10 levels (in $\mu g/m^3$) at six sites across Utopia on a daily basis for several years. They now ask you to perform an analysis which explores how the six sites compare in terms of the recorded PM10 levels.
The society is further interested in identifying factors which may explain differences across the sites. They are aware that their current data does not allow them to make such conclusions. You are thus asked to suggest other data or information the society may want to collect so that they can address this question in the future. **[15 marks]**

(2) The government of Utopia is in the process of drafting a new strategy paper on energy production. This initiative is driven by the government's recognition that a transition to net zero is essential to keep the rise in global surface temperature to well below 2°C above pre-industrial levels. The Department of Energy of Utopia has provided you with data on energy consumption, efficiency and production for the past years. They ask you to address the following two questions using the data provided:

(a) Which features regarding energy consumption and efficiency does the government of Utopia need to take into account when designing their new energy strategy? **[15 marks]**

(b) Energy in Utopia is produced using four sources: several solar farms, wind farms, coal-fired power stations and a nuclear power station. Explore how much renewable energy sources, solar and wind, have been contributing to the energy production.
The current capacity for energy production using coal is 500,000 MWh per day. An advisor to the government of Utopia suggests to reduce this capacity by 40% and to close the nuclear power station within the next two years. The latter is motivated by the country's difficulty to find a secure ultimate disposal place for the nuclear waste. What would be the impact of this plan on the country's capacity to meet its energy demand? Clearly state the assumptions underlying your approach.
Do you have any suggestions on how this plan may be improved? In your discussion, please take into account that Utopia is unable to import energy from abroad and to immediately build new solar and wind farms. **[15 marks]**

Data Descriptions

You have been given access to three data sets from the country of Utopia:

PM10.csv: Observations of PM10 collected by the society "Clean Air for Utopia" at six different sites.

Date: Time the observation was made

Site1: Recorded PM10 at the first site (in $\mu g/m^3$)

Site2: Recorded PM10 at the second site (in $\mu g/m^3$)

Site3: Recorded PM10 at the third site (in $\mu g/m^3$)

Site4: Recorded PM10 at the fourth site (in $\mu g/m^3$)

Site5: Recorded PM10 at the fifth site (in $\mu g/m^3$)

Site6: Recorded PM10 at the sixth site (in $\mu g/m^3$)

Utopia_Energy_Consumption.csv: Energy consumed by the country of Utopia over the period 2012-2024, measured at an hourly resolution.

Date: Date and time of the observation.

Demand: Amount of energy (in MWh) which was consumed by the population of Utopia in the time from **Date**-30 minutes until **Date**+30 minutes.

Utopia_Energy_Efficiency.csv: Energy efficiency ratings for 49530 properties in Utopia built before 2019.

Year: Year the property was built

Type: Type of property ('Flat', 'Terraced', 'Semi-detached' or 'Detached')

ERC2014: Energy efficiency rating for 2014 ('A' = 'Efficient', 'B' = 'Partly Efficient', 'C'='Inefficient')

ERC2018: Energy efficiency rating for 2018 ('A' = 'Efficient', 'B' = 'Partly Efficient', 'C'='Inefficient')

Utopia_Energy_Production.csv: Energy produced by the country of Utopia over the period 2012-2024, measured at a daily resolution.

Day: Date of the observation.

Solar: Amount of energy produced by solar farms on the day (in MWh)

Wind: Amount of energy produced by wind farms on the day (in MWh)

Coal: Amount of energy produced by coal-fired power stations (in MWh)

Nuclear: Amount of energy produced by the nuclear power station (in MWh)