Research Skills: Programming with R

Final Assignment: Group Project

Task Description

In this assignment, you must use R for a small research project. Find an existing data set, ask a research question about it, and use a novel R package to obtain insights into it. This novel R package might focus on data manipulation, visualisation, or modeling. Your final submission should consist of:

1. at least one data file
2. a least one R script containing all your code
3. a short PDF research report, containing:
    - a description of the origin of the data
    - one or more research questions
    - a description of how different R package(s) were used in the project
    - an overview of the contents of the data, using tables and / or plots
    - visualisations or model results that answer the research question(s)
    - some interpretations of these visualisations or model results
    - a brief conclusion

Group Work & Process

You will complete this assignment in groups of 5, which will be randomly assigned via Blackboard. This simulates a real workplace situation, and minimizes the time spent on group formation. Only students who submit the first worksheet will be enrolled into groups, and this will occur on Wednesday, October 31st, 3:00 PM.

The deadline for submission of all materials is **Wednesday, January 23rd, 5:00 PM**. Only one student per group should submit, and you will all receive the same grade. You should expect to spend about 20 hours per group member working on this assignment. All group members should contribute to writing code.

If group members leave the course, and your final group size shrinks, this is not necessarily a problem; as detailed below, the report requirements will be adjusted accordingly. If your group size shrinks to two or fewer, if there are other teamwork problems, or you missed group formation altogether, please contact us.

Product Requirements

Your data file should be:

1. included with your submission
2. OR, if it is very large: downloadable via a stable link (i.e., don't send it via WeTransfer)

Your R code should be:

1. complete and well-organized, divided into clear sections
2. easy to directly match up with your report
3. labelled with who wrote what section
4. labelled with your group name, names & u-numbers in every file

Your PDF research report should include:

1. a well-written report corresponding to the Task Description
2. (N - 1) visualisations
3. N * 250 words (+/- 10%)
4. your group name, names & u-numbers at the top of the report

where N is your group size.

Grading

The group project makes up 30% of your final grade, and it will be graded on five aspects: Efficiency & Correctness (2 points), Difficulty (3 points), Presentation (3 points), Purpose (1 point) and Requirements (1 point). For each aspect, we'll consider the following questions when deciding how many points to award:

For *Efficiency & Correctness*: Does your code do what you say it does, in a reasonable number of lines? That is, don't write out by hand what you could solve more efficiently by intelligently using functions.

For *Difficulty*: How ambitious are your data manipulations and chosen R package? The more processing steps, and the further away from topics presented in class, the better; two different R packages are recommended for full points.

For *Presentation*: How well-written are your report & code? Are your visualisations visually attractive and well-labelled, with clear captions? Does your code adhere to proper code style?

For *Purpose*: How novel and relevant is your research question, and can you actually answer your research question with the data set and methods chosen? (Note that this makes up only 10% of your grade, so don't optimise for this).

For *Requirements*: Does your submission adhere to the Product Requirements given above? (This should be a free point! Note that if you depart from the Requirements too much, your project may be returned, ungraded.)

Feedback & Re-Take

You will receive scores and brief feedback for all five of the aspects mentioned above. Should your group fail to hand in by the deadline date, or not pass on the first attempt, you can hand-in (an improved version) of your project by Wednesday, February 26th, 5:00 PM. The maximum grade will then be a 6.

Planning

Your planning is up to you, as long as the final product is ready by the deadline. We recommend that you contact each other via Blackboard as soon as groups are formed, and that you start looking for a suitable data set right away. You should have sufficient skills to start exploring and preparing the data set after completing Assignment 1.

Learning Goals

The main purpose of this group project is to test the following course goals: (i) find, install and use novel R packages; (ii) judge and improve the efficiency of code; (iii) combine these skills to solve novel problems.

Resources

Feel free to contact us before starting your project to discuss your ideas.

Your project should make extensive use of at least one R package not covered in class. Blackboard's Discussion Forum lists a few options; you may choose others if you wish. Not all R packages are suitable for all kinds of data; make sure your data, research question, and novel R package are all aligned.

For data sets, you can check e.g. the code competition website www.kaggle.com, the machine learning repository at UC Irvine, http://archive.ics.uci.edu/ml/, and the open data platforms of New York, https://nycopendata.socrata.com, London, https://data.london.gov.uk, and the United Nations, http://data.un.org.

To get a better idea of what's expected, you can check out two example projects on Blackboard, under Course Documents > Example Assignments; the code, report and feedback are all provided, though not the data sets. Do note that last year's requirements were slightly different.