Final assignment (70% of total mark)

Significance testing: continuous and categorical variables.
Correlation and linear regression.

Word limit: 3,500 words maximum, excluding the R script.

Deadline: Monday, April 24, 2023, 12pm

How to submit: This assignment must be submitted online via the Assignment area of our Moodle site. Please do not submit work by e-mail. Each copy must have a completed Coversheet for Coursework attached to it. The coversheet can be downloaded from our Moodle page.

When completing the assignment, please remember that…

- you should answer all of the questions and all the parts of each question;
- your assignment submission must include a complete log of your R session, including your R script with the commands and any graphical outputs produced. The log and the plots don't count towards the word limit;
- you do not need to write a lot in response to each question, but you should present your results, etc., in a neat, orderly, and professional manner, as if you were including them in a formal paper (e.g., a report or a dissertation). That is, do not just dump raw R outputs into the main text of your assignment;
- the word-limit for this assignment is a maximum upper limit only. You must be thorough and precise to do well, but you need not "pad out" your assignment to get anywhere close to this limit.

Are you ready? Then you can start the assignment on the next page! 👇

Question 1

The file undergrads.csv (available on Moodle) contains information regarding two groups of universities in the UK: the Russell Group and the Cathedrals Group. The information relates to the academic year 2020-2021 (source: Higher Education Statistics Agency, HESA).

The four variables included in the data set are:

- ukpc – the percentage of registered undergraduates who are domiciled in the UK
- eupc – the percentage of registered undergraduates who are domiciled in EU countries
- noneupc – the percentage of registered undergraduates who are domiciled in non-EU countries (other than the UK)
- fempc – the percentage of registered undergraduates who are female

The group to which each university belongs is coded as either russ or cath in the additional variable "group".

Conduct a statistical comparison of the two groups of universities in relation to each of the four variables. Write a very brief report on what kinds of analyses you conducted and what you discovered. Remember to justify your choice of method(s) for the data analysis.

Question 2

Table 1 contains information about the results of random mandatory drug testing in English prisons (source: HMPPS Annual Digest 2019-2020). It shows the numbers of positive tests for different types of drugs in the years ending March 2018 and March 2020 respectively.

Table 1. Mandatory random drug testing in English prisons: counts of positive tests by type of drug.

	2018	**2020**
Amphetamines	62	122
Barbiturates	2	3
Benzodiazepines	320	217
Buprenorphine	1063	779
Cannabis	3067	3082
Cocaine	259	259
Methadone	325	375
Opiates	1257	1110
Psychoactive Substances	6636	2203

Complete a statistical analysis to test whether the distributions of positive tests are different between the two years. Write a very brief report on what kinds of analyses you undertook and what you discovered. Ensure that you justify your choice of method(s) for the data analysis. If you find a statistically significant difference, comment on the size of – and the main contributors to – the overall effect.

Question 3

The data set usrelig.csv, available on Moodle, contains a number of social indicators, for the years 2008-2009, relating to the 50 states of the USA, plus the District of Columbia (Source: US Census Bureau, Statistical Abstracts of the United States). Data concerning the proportion of women smokers in each state are also included (source: Behavioral Risk Factor Surveillance System, 2009), as are two estimates of religiosity and religious affiliation (source: Pew Research Center, 2014; data for 2008-2009 were not available).

The following variables are included in the file:

- vrelig - the percentage of people in the state who described themselves as "highly religious"
- cathpc - the estimated percentage of Roman Catholics in the state's population
- abort - the abortion rate for the state (per thousand women)
- femsmoke - the estimated proportion of female smokers in the state (as a percentage of the state's female population)
- pov - the percentage of individuals in the state who live below the poverty level
- hsch - the percentage of the state's population who have graduated from high school
- univ - the percentage of graduates in the state's population (i.e. those with at least a bachelor's degree)
- urban - the percentage of the state's population which lives in urban (as opposed to rural) areas

Read the data set into R and use an appropriate method (or methods) to investigate a possible relationship between the percentage of "highly religious" people (as your outcome variable) and the percentage of graduates in the state's population (as your predictor variable).

In your report, make sure that you describe clearly what kinds of analyses you undertook and what you found out. Ensure that you justify fully your choice of method(s) for the data analysis. When performing any regression analyses, ensure that you also undertake and report **all** of the necessary diagnostics. Where appropriate, make sure that you also plot your regression model(s) and write down the model(s) in the form of an equation that we could potentially use to make predictions manually.

Question 4

Table 3 shows the sizes of the male and female prison populations in England in two corresponding weeks of 2021 and 2022 (source: HMPPS weekly reports).

Table 3. English prisons: numbers of prisoners by gender in corresponding weeks of 2020 and 2022.

	28-Feb-20	25-Feb-22
Female	3721	3202
Male	80147	76522

Undertake a statistical analysis to test whether there is any association between year and prisoner gender. Write a very brief report on what kinds of analyses you undertook and what you found out.

Ensure that you justify your choice of method(s) for the data analysis. If you do find a statistically significant association, ensure that you also comment on the effect size.