# MT5758 Assignment 1

**Due:** Friday 10 February 2023 (23:59 GMT)

You will be working on a project throughout the term, to get hands-on experience working with and analysing multivariate data sets. The project is broken into three assignments: a group proposal (10% of total grade), an individual outline (10%), and an individual report (30%). A document with a list of students in each group is available on Moodle, under the "Project" section.

### Instructions

For the first assignment, each group needs to find a multivariate data set, which will be the basis for your project. Then, as a group, you should answer the following questions within 1-1.5 pages (with 1-inch margins, font size 11pt). You can use a simple "Q & A" format, i.e., a list of each question followed by its answer. Although formatting is not particularly important for this assignment, clarity of writing and presentation is important.

- **What is your data set?** Provide some background about the origin of the data to motivate the analysis. Describe its dimensions, what the variables are (including type, e.g., continuous or categorical), what the data units correspond to.
- **What makes this data set specifically an interesting multivariate data set?**
- **Describe one or two possible directions of analysis.** You do not need to describe a method of analysis, but rather suggest interesting questions that could be explored. This will not bind you for future assignments.

You should submit one proposal per group on MMS, in PDF format. (It is up to you which group member does this.)

### Remarks

- Although you do not need to choose a method of analysis at this stage, it is helpful to be aware of the types of questions that can be addressed by multivariate methods. As mentioned during the first lecture, we will focus on two particular problems in this course:
  1. Can we reduce the dimensionality of a multivariate data set, e.g., for visualisation purposes or to save disk space? These methods are most useful when the observed variables are correlated.
  2. Can we group data units into clusters based on similarity?
- You should not use a data set that has previously been analysed using similar methods (e.g., from a textbook or research paper). You can use public data repositories, such as Kaggle or the UCI machine learning repository, as long as you do not follow someone else's analysis. If you do not know where to look for a data set, or if you have doubts about whether you can use a specific data set, you can discuss this with the module coordinator.