Coursework for M3S20 Introduction to Statistical Learning

This coursework counts for 10% of the overall mark for the module.
This coursework should take you between four and six hours to complete.
Date Coursework Set: 4th February 2020.
HAND-IN DEADLINE: 18th February 2020, by 5pm.
HAND-IN METHOD: Electronically, via Blackboard at the CW1 Turnitin item. Not by paper or email, please. If you have a special need for a different route, or require the coursework in a different format, please contact me.

Task.
You should find a suitable data set for analysis by distance-based methods. You should set out the reasons for the choice of your data set and what is of interest in your data set that is intended to be revealed by distance-based methods. You should analyse your data set using distance-based methods in R and then write a report about what you did and what your conclusions were. You should then submit your report using the Blackboard Online Learning Environment by the hand-in deadline. You should imagine you are in a working environment and writing your report for your technically knowledgable manager, who has to communicate/sell your conclusions to more senior management and stakeholders.

YEAR 4 students only: Those students taking the unit as a year 4 unit should undertake the following extra component. Using your set of data, formulate and carry out a separate statistical analysis on your data — either on the distances/dissimilarities from the main part of the project, or on the resulting configuration. You have an extra 0.5 pages to write up this analysis and your conclusions, but the overall additional figure/table count remains at ten.

Advice on how to undertake the task and associated conditions/rules follow:

1. Your data set can originate from any source. For example, this could be the Internet, from an academic paper or book, or data that you have collected yourself. Your report should make it clear where the data originates from and provide appropriate referencing.

2. Reasons for the choice of data set include: the data is of current topical interest, the data is related to an important subject, the outcomes of the analysis are particularly interesting or revealing, the data relate to your hobbies or interests but there has to be reason for using distance-based methods (e.g. a natural configuration does not exist, or cannot be found, or it is very easy to compute distances). This list of reasons is not exhaustive and you might have others.

3. You should ensure that you use multidimensional scaling as one of your analyses and explain what you are using and why.

4. You are NOT permitted to use any dataset that is built into R or one of its packages.

5. Your report should be submitted as a PDF file. The written content can NOT exceed two A4 pages using no less than a 10-point font. Your written report can be supplemented by up to ten additional figures or numerical tables with their associated captions (you do not have to include ten, but this is the maximum). You can use any text processing system to produce your report (e.g. Microsoft WORD or LaTeX), but the submitted document has to be a *single* PDF file. You

are permitted to produce dynamic or Shiny-type graphics online, but this is optional. If you do, each online figure/table counts to one of your allotted figures/tables. Provide the link to the dynamic graphic in your written report than can be clicked on to access the graphic.

6. Your report should not reproduce detailed mathematical development of distance-based methods as presented in lecture notes. You only have two pages to write about the data, your reasons, the analysis and reasons for using that analysis, the results, and conclusions..

7. Amongst other things, marks will be available for: (i) an interesting data set and good reasons for analysis using distance-based measures; (ii) a competently executed analysis of the data; (iii) a clear and cogent well-written report that includes a description of the results and conclusions; (iv) clear and informative figures/tables where necessary.