CS7DS3 Main Assignment
Advice and hints

Audience

- 8 page report analysing wine review data
- Intended audience should be two people
  - **Non-technical:** e.g., wine seller, who would like to use your analysis in **qualitative** way (e.g., which kinds of wine are usually good value for money?)
  - **Technical:** e.g., a line manager/supervisor/team leader. Someone who wants to make sure your analysis is correct. (e.g., did you check your MCMC output for convergence?)

Data

- Data set has interesting characteristics:
  - Use either `points` or `superior_rating` (not both)
  - `description` is free text – fine to ignore this but may be of interest to some in class
  - `variety` – consider grouping by this variable
  - `price` variable is skewed – worth investigating

Language and models used

- Use any language you like
  - i.e., Python is fine for your analysis instead of R if you really prefer it.
  - But I strongly recommend using the case studies as a template to follow.
  - I don’t know how easy it is to fit the most advanced models we used in class in Python – so be careful.

Language and models used

- You can use other methods as a **complementary** analysis, if you like, but this is not needed.
  - E.g., t-SNE for visualisation, NLP methods for description variable
  - Alternative classification methods to evaluate performance
  - Remember, predictive performance is only one of several things I will be interested in here, and I'm not that interested in it

Structure

- I recommend a basic structure as follows
  - Executive summary/objective statement followed by central figure
  - Introduction
  - Data description
  - Analysis (maybe further subsections)
  - Conclusions
    - Summarise results
    - Overall evaluation – e.g., were methods successful?
    - Further recommendations – would other methods/approaches/different data be more suitable?

Evaluation

- Models should be clearly explained.
  - N.B., summary statistics are not models
- Choices and decisions should be discussed and justified
  - N.B., “the [Python] output selected these variables as being important” is not analysis

General advice

- Start small and work your way up
  - Pick a small number of variables to start with and check your interpretation
  - Run a simple model before you run a complex one
    - Simple models explained well are better than complex ones that I can’t understand

General advice

- Use graphs where you can
  - Usually easier to interpret than tables
  - Sense check – does model output match data interpretation?
  - Can you visualise interesting features of data/model
- Don’t report redundant information
  - e.g., parameter estimates to 6 decimal places.
  - If you fit two models and one is better than the other, spend more time discussing the better model.

General advice

- Be creative:
  - Central figure can be composite of several figures/diagrams. See e.g., https://www.scss.tcd.ie/arthur.white/Teaching/STU33011/cluster_poster_dublin_nice_visualisation.pdf
  - What model you specify
  - Variable transformations and summaries
  - Use of complementary skills, etc.
  - Use of model **output**

Finally

- Use the discussion board and keep in touch
  - The earlier you post the quicker I can reply

Good luck