Grade Elements

- 20% of your final grade in the course.
- 15% written report (8–12 pages, not including R script to be passed in separately, more details below)
- 4% oral presentation (3–5 minute PDF)
  - 2% clarity/organization of presentation,
  - 1% accuracy of analyses/interpretation (proper usage),
  - 1% within time-limits (3-5 minutes)
  - Attendance, attentiveness, and respect is expected during your classmates' presentations. I reserve the right to penalize your presentation score if I'm unsatisfied in this regard.
- 1% Group Member Contributions assignment (on Canvas). These ratings well help me adjust grades in cases where a group member did not contribute to the project. You will give ratings for everyone in your group (including yourself, and including if you are the solo member of the project). These will remain strictly confidential. There will be no reference to reduced grades when I provide rationale for grades given, except the extreme circumstance where it is clear that an individual has provided NO contributions to the group project (requiring me to fail them on this grade item). Details on how to provide ratings will be given on Canvas, and discussed in class.

Due Dates

- By March 1: Get the OK on a data set — probably 10+ variables (4+ numeric, categorical and numeric response variable ideal), 200+ observations. First come, first served! Each group will be required to use a unique data set...
- By March 8: meet with me at some point to discuss progress and show some preliminary results in R (emailed screenshots are acceptable too):
  - Show that you can read in the data, and have been able to perform some of the analyses we have discussed in class
  - Not necessary to formally interpret the results for me.
  - This is to avoid mayhem closer to the deadline ("I'm getting errors but the project is due in 2 days!")
- March 31: Presentation should be emailed to me in PDF form. (If using powerpoint, it should be easy to export as PDF...you will lose any animated slide moves, so don't bother with those!)
- April 4: Written report due, R script due, Group member contributions due.
- Oral presentations (3-5 minutes each) are currently TBD (either in lab or in class).

Report Rubric

This is a guideline of what I will be using to mark your projects. Note that ratings out of 5 do not indicate that it is worth 5% of the written report grade.

All yes's and 5's would give some form of A, all 4's B, 3's C, 2's D, 1's F. Any "no" will incur a penalty (depends on the type and level of infraction).

1=Poor, 2=Unsatisfactory, 3=Satisfactory, 4=Good, 5=Outstanding

Met deadlines	No		Yes		
Meets length requirements	No		Yes		
Includes all required sections (as specified in this handout)	No		Yes		
Layout and organization	1	2	3	4	5
Clarity of writing (style/spelling/grammar)	1	2	3	4	5
Thoroughness of application (eg, uses enough analyses?)	1	2	3	4	5
Accuracy of application (eg, analyses correctly coded?)	1	2	3	4	5
Description of methods	1	2	3	4	5
Interpretation of results	1	2	3	4	5
Clarity/organization of R script	1	2	3	4	5

Report Details:

- Neatness of document counts, as does grammar/spelling. Communication is a key element in any mathematical/statistical job. It should be in a 'formal' tone (for instance, the sentence "I found this data..." should rather be "The data was found...").

- It should be aimed at a statistically competent audience, but do not assume that the reader has knowledge of all the specific analyses you are applying. The idea is for you to showcase your understanding of the methods.

- Report should include at least the following sections:

  - Introduction - A paragraph or two on what you are investigating (and why).
  - Data - Introduce the data set. How did you find it (citations necessary)? What information does it contain? Some preliminary descriptive statistics and plots would be appropriate here as well.
  - Methodology and Application - Explore the data set and explain the techniques you are using (pretend that it is written for someone with a decent statistical background, but no expertise in the analyses you are using). Point out interesting results.
  - Conclusion - Review any interesting results you found in a broader context. What might these imply? If no particularly interesting results were found, provide potential explanations why (for instance, confounding variables? noisy data?).
  - References - Plagiarism is a serious offence. Any quotes, or paraphrasing ideas from other authors need to be referenced appropriately (reference style is up to you, but should be consistent throughout the paper). How you obtained the data should be included (website of origin, or book/journal article that it appears in). If you're unsure if something needs to be referenced, ask me.

- Your R script should be provided to me via email at the time of submission of the written report. I should be able to run it on my computer (so the script should include how to load in the data). The data may be attached as a .csv file if it is externally sourced.

Analyses

I would expect a minimum of one analysis from each of the three major statistical learning areas (regression, classification, and clustering) and a minimum of 7 total from the list that follows. Just like the bias-variance trade-off that is a cornerstone of this course, there is a trade-off between the quantity and quality of analyses! It may be difficult to fit every analysis (including plots, tables, etc), and a proper summarization of the methods and results, in a maximum 12 page report. You are better off with fewer, well-done analyses. Remember, you are showcasing your knowledge! Just running all the analyses by changing my R code and throwing them into a document will NOT lead to a satisfactory grade.

Exceptions to the minimums provided in the previous paragraph can be made in extreme circumstances, but keep in mind (for example) that if you only have one potential continuous response variable in your data, it is trivial to turn that into a binary response which can then be used for classification!

Here is a list of analyses from the course. * indicates that I expect we will cover them in time for you to apply to your project data.

- Multiple Linear Regression (including interactions, possibly mixed variable types)
- Variable Selection for Linear Regression
- K-Nearest Neighbours Regression
- K-Nearest Neighbours Classification
- *Linear Discriminant Analysis
- *Quadratic Discriminant Analysis
- *Hierarchical Clustering
- *K-means Clustering
- *Logistic Regression
- *Regression Trees
- *Classification Trees
- *Random Forests/Bagging
- *Principal Components Analysis
- *Factor Analysis
- *LASSO
- *Neural Networks
- *Mixture Models
- *Non-negative matrix factorization

In most cases, cross-validation should be used for model or tuning parameter selection, but this is a 'tool' rather than analysis. Same goes for bootstrap (except in the case of random forests/bagging, which are separate analyses based on that tool) and validation sets. If CV is never used, expect major deductions!