Final Project

1 Instructions

- The final project is due on **Wednesday, December 6th at 11:00pm**.

  This is a group project, **late or missed submission is strictly NOT permitted and will be awarded a grade zero**. Alternative arrangement will only be considered under exceptional circumstances, and has to be made with the instructor **prior to the deadline** and at the discretion of the instructor.

- You may work in teams of up to 3 students, and each team will work on a designated data set. Submit one project per team.

- Each project consists of a computer-typed report **strictly between 6-8 pages including figures**, but excluding a mandatory Appendix containing (but not limited to) all `R` code.

- Reports must be submitted in PDF format electronically via LEARN.

  Login to LEARN. At the top of the screen, click

  `Assessments` > `Dropbox` > `GroupX` : `Project` > `Add a File`

  The names of all collaborators must be written on your report

2 Proposed Outline of Report

1. Summary (Mandatory)
   A maximum of 200 words describing the objective of the report, an overview of the statistical analysis, and a summary of the main results.

2. Model selection
   Use linear regression methods, consider several models for response as a function of other variables in the dataset.

   - Your analysis must contain at least one automated model selection.
   - Discuss any issues encountered during model fitting and how you addressed them, e.g. non-linearity issues, heteroscedasticity issues, multicollinearity issues, etc.

- Narrow down two candidate models for closer inspection.

3. Model diagnostics
   Perform an in-depth comparison of the two candidate models you have proposed by examining the following diagnostics:

   - Different types of residual plots.
   - Leverage and Influence measures.
   - Cross validation.
   - Make your predictions on the regular scale.

   Analyze these diagnostics, and retain one final model. Display its parameter estimates and confidence intervals.

4. Discussion
   Report your findings in the context of the data. For example:

   - What are the most important factors affecting the response? Are there any coefficients with high p-values retained in the final model? If so, why?
   - Are there any outlying observations that might be appropriate to remove?
   - Are any of the regression assumptions of the final model violated? If so, which ones? What are the possible deficiencies of the final model. How the model can be improved? (you may use materials outside of this course)

3 Grading

You will be graded for content (50%) and presentation (50%) of the report.

Content:

- Technical challenge.
  Appropriateness/originality of variable transformations to improve fit/tell a better story.
- Correct and efficient programming.
- Correct and insightful interpretation of the results.

- Justification of subjective decisions.

Presentation:

- Organization of information, overall legibility.
  Present only the most relevant models and output, optionally including further analyses in the Appendix.
- Clarity of explanations.
  Use full sentences. Avoid using abbreviations such as Demo when giving explanations.
- Properly commented `R` code.
  A suggestion is to divide the Appendix into clearly labeled blocks of code, each starting with a description of what it does and where to find it in the report.
- Properly labelled figures, succinctly formatted regression output.
  Include figure captions or titles, properly labelled axes. Do not waste space by displaying the entire output of a summary command.