Statistical Learning Project

This project may be individual or as a group of no more than 5. Pick a real data set for which you believe there are interesting questions to answer. You will then try out all the different statistical learning approaches that we have covered in this course to try to find the best way to answer these questions.

Deliverables
This project includes three deliverables:

1. A proposal for the project- one page long
   (a) Description of the problem
   (b) Description of the dataset (dimensions, names of variables with their description)
   (c) Supervised or Unsupervised?
   (d) Regression or classification?
   (e) Comments and/ or concerns?
   (f) Proposal is due by midnight Fri. Oct. 25. Approvals are first come first served.
2. Project presentation
   (a) Description of the data and the question/s that you are interested in answering.
   (b) Review of some of the approaches that you tried or thought about trying.
   (c) Summary of the final approach you used and why you chose that approach.
   (d) Summary of the results.
   (e) Conclusions.
   (f) Make sure you articulate your work clearly and concisely. Points will be given for the formulation of the question of interest, the approach or approaches you used to solve those questions, the reason you chose a particular approach, and the conclusions you were able to draw.
3. A final report to be handed in. The report will contain a summary of the material covered in the presentation (maximum 3 pages). The first page should be a summary. The report must also include the slides from the presentation and a technical appendix, which should include your R/Python code (maximum 10 pages).

- Data Repositories
  1. Open Gov. Data: Open Government - Data.gov
     https://www.data.gov/open-gov/, www.data.gov.uk,
  2. KDD Nuggets: http://www.kdnuggets.com/datasets/
  3. UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/
  4. StatLib: http://lib.stat.cmu.edu/datasets/
  5. TwitteR: http://cran.r-project.org/web/packages/twitteR/index.html
  6. rfigshare: http://figshare.com
  7. r-interface for figshare: http://cran.r-project.org/web/packages/rfigshare/index.html
  8. Note: Cutting and pasting the URLS may not work in some instances since the typset PDF document may have hidden characters (ligatures). If cut and paste does not work, you may have to manually type in the URL.
- Suggestions
  You should be able to reuse R code from the class Labs.