DSO 530: Applied Modern Statistical Learning Methods
Assignment 6 (Due 11/13/2014)

Guidelines for assignment submission:

1. Type each question before you answer it, and provide a clear separation between each part.
2. All relevant computer output should be provided unless noted otherwise.
3. Print your homework, and submit it at the beginning of the class. Make sure that it is stapled, and your name is typed on it.
4. Attach your R code as an Appendix. Make sure to provide comments on what your code is doing. Keep it clean and clear!

5. Note the main aim of this homework is to get practice doing linear regression in R. Hence you shouldn't restrict yourself to only doing specifically what is asked. Anything else you might want to do to build a better linear regression model would be welcome. Questions 1 and 2 should either be answered using the auto data set from lab 1 or, if you are feeling bored with this data, using a data set you have gathered. Some good sources for interesting data sets are
   http://www.econ-datalinks.org/search.html
   http://fisher.osu.edu/cgi-bin/DB_Search/db_search.cgi?setup_file=finance.setup.cgi
   http://fisher.osu.edu/fin/fdf/osudata.htm
   http://www.census.gov/epcd/www/recent.htm
   http://www.bized.ac.uk/dataserv/freedata.htm

   Warning: These websites are fascinating, with thousands of possible data sets to explore. You may find yourself ignoring your loved ones and your other classes just so that you can sneak back and spend more time exploring. This is unfortunately one of the dangers of taking a statistics class and is unavoidable! Seriously, real data sets often have real problems (that have nothing to do with R) associated with them. If you run into problems you may find it easier to work with the auto data.

DSO 530: Applied Modern Statistical Learning Methods
Assignment 6 (Due 11/13/2014)

Look for the fgl dataset in MASS package.

1. Describe the dataset.
2. Compare the accuracy of the following models to predict the response variable type in the fgl dataset using all predictor variables.
    a. LDA
    b. Classification Trees
    c. Bagging
    d. Random Forest.

    Which one was the winner? Report the misclassification errors for all of the 4 models. Use set.seed(1) when needed.