Analytics Foundations: Lab 6

1. Use the same cars2010 dataset you have used in previous labs. To obtain this data, submit the following code:

```
library(AppliedPredictiveModeling)

data(FuelEconomy)
```

   This dataset has variables pertaining to fuel economy of various cars. **Do not** create a training and test set. Just use the whole cars2010 dataset for the following analysis. The cars2011 and cars2012 datasets will be used at later time periods.

   Perform the following analysis:

   a. Run a regression predicting the **FE** variable using all the remaining variables. Some of these predictor variables are coded as numeric, but should be treated as categorical. The only numeric variables in your dataset should be **EngDispl**. All remaining variables are categorical.
      a. Perform a Global F-test. What is your conclusion?
      b. What percent of variation in fuel economy (**FE**) is explained by these 13 variables?

   b. Trying to evaluate categorical variables in traditional linear regression output can be difficult because the p-values are for each categorical dummy variable. To evaluate the inclusion of a variable as a whole, you need a global p-value for each categorical variable.
      a. Use the `anova` function in R on your linear regression object to get the p-values for each categorical variable.
      b. Which of the variables has the highest p-value?

   c. Rerun the preceding model, but remove the variable with the highest p-value that you found with the `anova` function. Compare the output with the preceding model.
      a. Did the p-value for the model change notably?
      b. Did the R-square and adjusted R-square values change notably?
      c. Did the p-values on other variables change notably?

   d. Again, rerun the preceding model (from question c), but eliminate the variable with the highest p-value. Repeat this process of eliminating one variable at a time and rerunning the regression until you only have variables significant at the 0.008 level. Remember to run the model after EACH variable you remove as the p-value might change by removing a variable.
      a. Did the R-square and adjusted R-square values change notably?
      b. How many variables did you have left that were significant at the 0.008 level?