**Assignment 2: Multiple Regression Analysis**

**The Data Set**

The “Boston Housing” dataset recorded properties of 506 housing zones in the Greater Boston area. Typically one is interested in predicting MEDV (median home value) based on other attributes.

Here is the list of attribute information:

1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centers
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. LSTAT: % lower status of the population
13. MEDV: Median value of owner-occupied homes in $1000's

**Pre-Processing**

MEDV has a somewhat longish tail and is not so normally distributed, so we will take a log transform, (use LMEDV= log(MEDV) ), and then predict LMDEV instead. You should convince yourself that this is a better idea by looking at the histograms to assess normality; however no need to submit such plots.

**Questions:**

1. Please perform the multicollinearity diagnosis based on the VIF calculation results. Do we need to drop any variables that might have multicollinearity concerns?
2. Please run the linear regression analyses.
3. Use the stepwise model selection approach to determine the final model. Drop variables based on their significance.
4. Report summary output for ***each step***, including Analysis of Variance (ANOVA), R-Square, and Parameter Estimates.
5. How do you interpret the final model?

**Software Selection:**

*Option A: Excel*

* File🡪Options🡪Add-Ins🡪Analysis ToolPak
* Note: need to calculate VIF manually.

*Option B: SAS Enterprise Guide*

* Use Query Builder🡪Computed Columns🡪Advanced Expression🡪LOG functions for pre-processing
* Regression: Tasks🡪Regression🡪Linear Regression
  + In Statistics, you can select Diagnostics🡪Variance Inflation Values to automatically calculate VIP