STAT 1221/2220
Homework 17

Name: Stephen Szasko

Problems 1a, 1c, and 1f require the use of R. Be sure to install the regclass package before beginning this assignment. Upload the knitted R file with the completed problems.

1. The Boston housing dataset is a famous dataset compiled in 1978 based on data from the U.S. Census Service. Each of the 90 observations in this dataset represents a neighborhood in the Boston metropolitan area. The goal of the dataset is to describe median house price in the neighborhood (Y), denoted **PRICE** according to many different factors. Included in this dataset are the following predictors:

   - X_1: CRIME: Per capita crime rate by neighborhood
   - X_2: ZONE: Proportion of residential land zoned for lots over 25,000 square feet
   - X_3: INDUS: Proportion of non-retail business acres in the neighborhood
   - X_4: NOX: Nitric oxide concentration in parts per 10 million
   - X_5: RM: Average number of rooms per residence
   - X_6: AGE: Proportion of homes in the neighborhood build before 1940
   - X_7: DIS: Average distance to five Boston employment centers
   - X_8: RAD: Measure of accessibility to radial highways
   - X_9: TAX: Property tax rate per $10,000
   - X_{10}: PTRATIO: Student-teacher ratio for schools in the neighborhood
   - X_{11}: BLACK: Percentage of Black residents in the neighborhood
   - X_{12}: LSTAT: Percentage of residents living below the poverty level

   Read the CSV file **boston_housing.csv** into R and use it to answer the following questions.

   (a) Fit the model that uses all 12 predictors to describe median house price. Print both the model summary and variance inflation factors. This part requires the **regclass** package.

   (b) Classify the predictors into three groups: those for which collinearity is not a problem, those for which collinearity is slightly problematic, and those for which collinearity is a severe problem.

   No Problem: $X_1, X_2, X_5, X_{10}, X_{11} < 4$

   Slight Problem: $X_3, X_4, X_6, X_7, X_{12}$ between 4-10

   Severe Problem: $X_8, X_9 > 10$

   (c) Create a correlation matrix of all 13 variables. To make it easier to view, round the correlations to 3 decimal places. This can be done using the 'round' function. For example, if your data frame is called 'df', your code would be **round(cor(df), 3)**, where the '3' represents the number of decimal places you want to round each correlation to.

   (d) List all pairs of predictors for which the **magnitude** of the correlation exceeds 0.70 along with their correlation.

Variables	Correlation
Dis - age	-0.714
Indus - Tax	0.744
Indus - NOX	0.801
NOX - Age	0.713

Variables	Correlation
Age - LStat	0.704
Rad - Tax	0.957
NOx - Dis	-0.823
Dis - indus	-0.757