ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
SCHOOL OF BUSINESS
MSc IN BUSINESS ANALYTICS

STATISTICS FOR BUSINESS ANALYTICS I
Assignment 2

The data for this assignment are a random sample of 63 cases from the files of a big real estate agency in USA concerning house sales from February 15 to April 30, 1993. The data was collected from many cities (and corresponding local real estate agencies) and is used as a basis for the whole company. The variables in this datasets are:

1. PRICE = Selling prices (in hundreds$)
2. SQFT = Square Feet of living space
3. AGE = Age of home (in years)
4. FEATS = Number out of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access
5. NE = Located in northeast sector of city (1) or not (0)
6. COR = Corner location (1) or not (0).

(I) Read the "usdata" dataset and use str() to understand its structure.

(II) Convert the variables PRICE, SQFT, AGE, FEATS to be numeric variables and NE, COR to be factors.

(III) Perform descriptive analysis and visualization for each variable to get an initial insight of what the data looks like. Comment on your findings.

(IV) Conduct pairwise comparisons between the variables in the dataset to investigate if there are any associations implied by the dataset.(Hint: Plot variables against one another and use correlation plots and measures for the numerical variables.). Comment on your findings.
Is there a linear relationship between PRICE and any of the variables in the dataset?

(V) Construct a model for the expected selling prices (PRICE) according to the remaining features.(hint: Conduct multiple regression having PRICE as a response and all the other variables as predictors). Does this linear model fit well to the data? (Hint: Comment on R^2 adj).

(VI) Find the best model for predicting the selling prices (PRICE). Select the appropriate features using stepwise methods. (Hint: Use Forward, Backward or Stepwise procedure according to AIC or BIC to choose which variables appear to be more significant for predicting selling PRICES).

(VII) Get the summary of your final model, (the model that you ended up having after conducting the stepwise procedure) and comment on the output. **Interpret the coefficients**. Comment on the significance of each coefficient and write down the mathematical formulation of the model (e.g PRICES = Intercept + coef1*Variable1 + coef2*Variable2 +…. + ε where $\varepsilon \sim N(0, ...)$). Should the intercept be excluded from our model?

(VIII) Check the assumptions of your final model. Are the assumptions satisfied? If not, what is the impact of the violation of the assumption not satisfied in terms of inference? What could someone do about it?

(IX) Conduct LASSO as a variable selection technique and compare the variables that you end up having using LASSO to the variables that you ended up having using stepwise methods in (VI). Are you getting the same results? Comment.