Data Exploration and Multiple Linear Regression (MLR) using R[1]

The "Boston Housing" data set, part of the MASS package, records properties of 506 housing zones in the Greater Boston area. For a description of the data (housing data and attribute information), visit https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names

Attribute Information:
Predictors:

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 centres**
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. B	**1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town**
13. LSTAT	**% lower status of the population**

Response:

14. MEDV **Median value of owner-occupied homes in $1000's**

1. Data Exploration:
    a. Check for missing values in the dataset.
    b. Plot the distribution of MEDV. What do you observe?
    c. Generate box-plots of the AGE (proportion of owner-occupied units built prior to 1940) and MEDV (median home value) attributes and identify the cutoff values for outliers.
    d. Generate a scatterplot of MEDV against AGE; comment on how inclusion of the outliers would affect a predictive model of median home value as a function of AGE.

2. Try to fit an MLR to this dataset, with MEDV as the dependent variable. MEDV has a somewhat long tail and is not so Gaussian-like, 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 and quantile plots to assess normality; however, no need to submit such plots). Keep the first 356 records as a training set (call it Bostrain) which you will use to fit the model; the remaining 150 will be used as a test set (Bostest). Use only LSTAT, RM, TAX, AGE and ZN as independent (predictor) variables and LMEDV as dependent (target) variable as follows when constructing a linear regression model:

$$LMEDV = \beta_0 + \beta_1 LSTAT + \beta_2 RM + \beta_3 TAX + \beta_4 AGE + \beta_5 ZN$$

   For this regression model, paste outputs for summary and ANOVA

[1] You must use R to run regression although use of other software is also encouraged for verification of your answers.

3. Do any variables have to be dropped because of multicollinearity? (Use VIF criteria to check for multicollinearity)

4. Report the coefficients obtained by your model. Would you drop any of the variables used in your model (based on the t-scores or p-values)?

5. Rerun your regression model after removing variables (if any) based on your analysis in the previous question. What is the value of R^2? What does it signify?

6. What is the overall F-statistic and the corresponding p-value of this final model? What does it signify?

7. Report the MSE obtained on Bostrain. How much does this increase when you score your model (i.e., predict) on Bostest?

8. (Bonus 1 point). Use the stepwise regression considering to reach your final model (LMEDV as dependent variable and all but MEDV as independent variables). Try different model section criteria (i.e., AIC, Cp, BIC, adj R^2) and see if you can come up with the same model even with the different criteria. Determine the best model if you get different models with different criteria? We will consider a model that gives the highest accuracy (in terms of MSE) in the test set as the best model.