COMP809 – Lab 3

This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. The data was originally published by Harrison, D. and Rubinfeld, D.L. '*Hedonic prices and the demand for clean air*', J. Environ. Economics & Management, vol.5, 81-102, 1978. There are **14** attributes which are:

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

Note: Variable #14 seems to be censored at 50.00 (corresponding to a median price of $50,000); Censoring is suggested by the fact that the highest median price of exactly $50,000 is reported in 16 cases, while 15 cases have prices between $40,000 and $50,000, with prices rounded to the nearest hundred. Harrison and Rubinfeld do not mention any censoring.

1. Build a simple linear regression to predict MEDV (house prices) using the RM (number of rooms) as a predictor.
    a. Check the assumptions.
    b. Interpret the estimated coefficient associated to RM.
2. Fit a multiple linear regression model to predict MEDV.
    a. Is there any categorical variable?
    b. Is multicollinearity present in the predictors? If so, how can you decrease its impact.
    c. Calculate the principal components and fit the linear model. Comment on it.
3. Train a multiple linear regression model to predict MEDV considering all the predictors.
    a. Fit the model using only 70% of the data. The remaining 30% will be used as a testing data set. Hint: use sklearn.model_selection.train_test_split function to generate the data sets.
    b. Interpret one of the estimated model coefficients and the R-squared.
    c. Compare the model to the one fitted in 2.c.

d. Evaluate the model performance on the testing dataset. For this, compare predicted values vs residuals and responses vs predicted values for both data sets.
e. Calculate the mean squared error (MSE) for both train and test set. Comment on it. Hint: use mean_squared_error function.