MIS 776  
Assignment 2 – Regression

Data Scenario

The three most important features impacting housing process are said to be “location, location, and location.” It is the characteristics of the neighborhood in which the house is located that will impact the house price as much as the characteristics of the house itself. In this assignment, we will test this concept by seeing if we can predict median home prices in various neighborhoods in Boston based on the characteristics of that neighborhood.

In this data set, we will attempt to predict the median house price (MEDV), which is a continuous variable, based on other continuous and categorical features of the neighborhood. The details of the data set are below. This data set is a modified version of the set 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.

Data

The target classification (output) column is *MEDV*. All other columns are potential predictors.

**CRIM: per capita crime rate by town   
ZN: proportion of residential land zoned for lots over 25,000 sq.ft.   
INDUS: proportion of non-retail business acres per town**

**RIVER: Yes if the tract is bordered by the Charles river, No of not.   
NOX: nitric oxides concentration (parts per 10 million)   
RM: average number of rooms per dwelling   
AGE: proportion of owner-occupied units built prior to 1940   
DIS: weighted distances to five Boston employment centres   
RAD: index of accessibility to radial highways   
TAX: full-value property-tax rate per $10,000   
PRATIO: pupil-teacher ratio by town   
LSTAT: % lower status of the population   
MEDV: Median value of owner-occupied homes in $1000's**

**Tasks** (Justify your answers by providing any screen shots or values you have observed that led to your conclusions.)

Submit answers to the following questions as well as the Jupyter Notebook that you used to get the answers.

1. Open the Housing.csv file in a text editor or Excel and evaluate the file. Make sure that you are comfortable with the various columns and their meanings are defined in the data structure outline above. Pay special attention to case, since Python is case sensitive. Close the file when you are done.
2. Create a new Jupyter Notebook and write the script to open the file into a data frame called *datHousing*. Do some exploratory analysis on the data frame to ensure that it was opened correctly and that everything looks reasonable. What is the mean value for MEDV? This is important, because the “simple model” for regression is to assume that all values for the dependent variable lie at the mean.
3. Create a dummy variable for RIVER called RIVERCODE where No is coded as 0 and Yes is coded as 1. Verify that this column has been added correctly by looking at the first few rows of your data frame. The mean value for RIVERCODE should be 0.069. Check this using the appropriate code.
4. Create a correlation table for all of the numeric columns in the data frame. This should be every column except for RIVER. Do you see any multicollinearity issues? Do you see any reason to try a Principal Component Analysis? Which predictors have the highest and lowest correlations with MEDV? Why is this important?
5. Create a correlation plot matrix using the same data as above. Focus on the relationships with MEDV. Do you see any potential issues such as non-linear relationships or heteroscedasticity?
6. Create a regression model predicting MEDV using all of the other numeric predictors. Remember to use RIVERCODE instead of RIVER.
7. Generate a summary of this model. Answer the following questions:
   1. What percentage of the variance in MEDV can be explained by this model?
   2. Assuming all other predictors are held constant, what is the effect does being by the river have on the median house price?
   3. Does the age of the home seem to matter?
8. What is the regression equation inferred by this model?
9. Using this model, what would we expect to be the median home price on a neighborhood given the following values:
   1. CRIM = 1.28955
   2. ZN = 0
   3. INDUS = 18
   4. RIVER = Yes
   5. NOX = 0.862
   6. RM = 5.22
   7. AGE = 65.2
   8. DIS = 2.05
   9. RAD = 5
   10. TAX = 399
   11. PRATIO = 13.9
   12. LSTAT = 18.3