**Problem Set 8 (due 11:59 pm, 21 December 2021)**

1. **Boston Housing Data**

This problem should be completed with Python.

Download the Boston Housing Dataset from:

<https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html>

There are 14 attributes. However, the 12th attribute is considered to be unethical and will not be used. (See the explanation in SK Learn.)

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.

CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)

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

PTRATIO - pupil-teacher ratio by town

~~B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town~~

LSTAT - % lower status of the population

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

1. Simple linear regression
2. Calculate the regression of MEDV using LSTAT as the input. Use LinearRegression in sklearn.linear.model. Provide the coefficients and

1. Plot the regression line together with the scattered data points.
2. Calculate the 95% confidence interval of the coefficients. Use t in scipy.stats. Note that t.ppf calculates the inverse cumulative distribution function.
3. If LSTAT = 4, 10, 14, what are the 95% confidence intervals of expected MEDV?
4. If LSTAT = 4, 10, 14, what are the 95% prediction intervals of MEDV?
5. Multiple linear regression
6. Calculate the regression of MEDV using LSTAT and AGE as inputs. Provide the regression coefficients.
7. Calculate