Open the file “Housing raw data.unsb”

The data set is described by the following variables:

Independent variables (X):

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 oxide 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 centers
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 [people of African American descent] by town
LSTAT: Percentage of lower status of the population

Response variable:

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

Procedure:

- Calculate the correlation matrix between all X-variables : Tasks – Analyze – Descriptive Statistics. In the project tree under results, click on Variable Correlations. Interpret this table with the description of the variables above in mind. Are the correlations as expected?
- Divide the 506 samples into training (2/3) and test (1/3). Create row sets in the Define Range editor (Edit – Define range). Hint: Click on Special intervals.
- Make an MLR model on the samples set “training” with the column sets “X-variables” and MedianValue(Y) as Y. Interpret the plots (ANOVA table, predicted vs. reference, residuals)
- Look close into the p-values, how are these related to the correlation table? You may also make a PCA model on the X-variables with weights = 1/Stdev.
- Plot the regression coefficients: Plot – Regression coefficients – Line. Change to bar plot with the icon  on the toolbar. Interpret this plot.
- Recalculate without some of the variables and compare the models. Mark variables and do Tasks – Recalculate - Without marked - Variables. Compare the results
- Now predict the test set: Tasks-Predict-Regression. Is the RMSE similar to the model on training set?