Missing Data Assignment

**Step 1: Baseline**

Using Sklearn get the Boston Housing dataset.

Fit a linear regressor to the data as a baseline. There is no need to do Cross-Validation. We are exploring the change in results

What is the loss and what are the goodness of fit parameters? This will be our baseline for comparison.

**Step 2: (repeated)**

Select between 1, 5 10, 20, 33, and 50% of your data on a single column (Completely at random), replace the present value with a NAN and then perform an imputation of that value.

In. each case perform a fit with the imputed data and compare the loss and goodness of fit to your baseline.

**Step 3:**

Take 2 different columns and create data “Missing at Random” when controlled for a third variable (i.e if Variable Z is > 30, than Variables X, Y are randomly missing). Make runs with 10%, 20% and 30% missing data imputed via your best guess. Repeat your fit and comparisons to the baseline.

**Step 4:**

Create a Missing Not at Random pattern in which 25% of the data is missing for a single column. Impute your data, fit the results and compare to a baseline.

**Step 5 (Extra Credit) (10 points):**

Using the MCMC method, and your data from step 4, What is the difference in performance between imputation via ‘guess’ (mean/median, etc) and MCMC.