**“Working with Missing Data”**

**Step 1:**

Use sklearn.datasets to get the Boston Housing dataset.

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

**Question 1:** What is the loss and what are the goodness of fit parameters? This will be our baseline for comparison.

**Step 2: (repeat for each percentage value below)**

Select **1%, 5% 10%, 20%, 33%, and 50%** of your data in a **single column** [*hold that column selection constant throughout all iterations*] (Completely at random), replace the original value with a NaN (i.e., “not a number” – ex., np.nan) and then perform an imputation for the missing values.

**Question 2:** In each case [1%, 5%, 10%, 20%, 33%, 50%] perform a fit with the imputed data and compare the loss and goodness of fit to your baseline. [**Note:** you should have (6) models to compare against your baseline at this point.]

**Step 3:** Take two columns and create data “Missing at Random” when controlled for a third variable (i.e., if Variable Z is > 30, then Variables X, Y are randomly missing). Use your preferred imputation method to fill in **10%, 20% and 30% of your missing data**.

**Question 3:** In each case [10%, 20%, 30%] perform a fit with the imputed data and compare the loss and goodness of fit to your baseline. [**Note:** you should have (9) models to compare against your baseline at this point.]

**Step 4:** Create a “Missing Not at Random” pattern in which **25%** of the data is missing for a single column.

**Question 4:** Perform a fit with the imputed data [25%] and compare the loss and goodness of fit to your baseline. [**Note:** you should have (10) models to compare against your baseline at this point.]

**Step 5:** Describe your imputation approach and summarize your findings. What impact did the missing data have on your baseline model’s performance?