Wrangling The Boston Dataset

Two approaches to getting the data for wrangling will be discussed and is documented in the jupyter notebook document on all the steps taken with this process. The first approach was to download the dataset into a text file. Delete the top lines of the dataset containing strings with Notepad ++ was done in attempting to convert all data entries to their appropriate floats. This approach proved to be unsuccessful because the shape of the chart could not be changed very easily. The second approach was to get the data using scikit learn, which proved to be successful since the shape was no longer a problem.

The first approach:

Pandas and numpy were imported so that I could convert the raw data to a pandas dataframe. Before opening it in Jupyter Notebook, the raw data was edited by deleting all the string above the actual data. The problem was when importing the data, there was only one column where there should have been thirteen. All the float entries that should be dispersed through out thirteen columns was all in one column instead. The shape of the data needed to be changed; more specifically the number of columns. It would require re-assigning all the columns. All lines would have to be iterated in the form of lists and all entries would need to be converted into strings and split by their respective tab separations so that only the floats are indexed. Only then could all the floats from each line be dispersed into their separate respective columns assigning each column to the corresponding entry by their indices. This would be a large mountain to climb. Other approaches to solving the problem regarding the number of columns could be used.

The second approach:

Further down the Jupyter Notebook document, I re-imported other modules like pandas and numpy along with importing other modules like sklearn. More specifically, Bunch was imported from sklearn.utils, datasets was imported from sklearn. For creation of strip plots and boxplots, seaborne was imported along with matplotlib.pyplot for creating the histograms. After successfully importing the data into a pandas dataframe, I renamed the columns. The index column can be left since each integer represents a random sub hurb in Boston.
Thirteen series were created; one for each column. These series were visualized with univariate plots namely the strip plot with jitter, the box plot and the histogram. These data were visualized this way along with verifying reasonable max/min values within each column and within their context. After visualizing the data, it was apparent that no outliers or missing values were present (since the info command on the data frame stated that there were no missing values and all entries were of one datatype; float). All the columns portrayed various obvious distributions like symmetrical/skewed normal distributions, exponential growth/decay and binary distributions. After determining these factors, the data has been cleaned and is now ready for analysis. Please refer to the associated Jupyter Notebook on how the data was wrangled for more detailed information.