DATA602: HW1/Review

My understanding of the problem:	The problem in looking at the dataset is to determine what predictor variables from a dataset can best-predict the housing price in Boston. These variables include: crime rate, residential zoning, size of industry, location, student teacher ratio, and others, and are analyzed to determine the target variable, median value of the house.

The approach:	**Linear Regression on Boston Housing:** On a high level, the method that the Agarwal uses is
	• Import necessary libraries
	  ○ Matplotlib, Seaborn for visualizing, Pandas/Numpy for data manipulation, and SKlearn for regression.
	• Preprocess the data- check to make sure there are few null values, and check distribution of the target variable (is it normal?).
	• Analyze and load the dataset by looking at the data and its descriptions- what does each variable mean? Pick target variable- Median housing price (target of analysis) and feature variables (other variables, predictors).
	• Analyze the dataset for correlations, multicollinearity, and look for the best predictor variables; graph predictive variables to look at correlation.
	• Prepare the data- split into train/test set with train_test_split, choose size of test, and random_state, which is allows for reproducible results (specifies seed value, gives same results every time, rather than random everytime)
	• Run sklearn linear_regression, which fits the model on the training set, then analyzes based on RMSE and R-squared.
	  ○ ***RMSE*** = *Root of Mean Squared Error, Each error per predicted vs. actual value is added, and the square root is taken of the average of error.*
	  ○ ***R-squared*** = *Total Sum of Squared Error divided by Total Sum of Squares*
	• Model yielded .66 on the training set, so 66% of the values could be explained by the model, OK performance.
	
	Class NB: Linear Regression:
	• In contrast, the class notebook calculates out and returns each linear regression value.
	• Created a fake dataset to show linear regression function.
	• Fit the regression with the linear regression function.
	
	Boston Dataset:
	• Same initial process as the article, but box plots each of the variables.
	• Selects all variables for running a linear regression instead of a few of them, fits a line, and predicts a value for price for 501.
	• Look at the different columns to determine the highest and lowest coefficient values (the correlation)== 'NOX' and 'RM'.

Ridge Regression	• Used for when data has multicollinearity (multiple strongly-correlated values), where least-squares is less useful. • Sometimes, models may overfit with regular linear regression. • Smaller weights can result in more stable/less likely to overfit.
Lessons Learned:	• Specifically from the dataset's correlations, it was interesting to look at a few of the predictive variables. Unfortunately, the LSTAT (lower income, see: https://opendata.stackexchange.com/questions/15740/what-does-lower-status-mean-in-boston-house-prices-dataset) had a strong negative correlation to housing prices; logically 'RM' or number of rooms, correlated strongly positively. • Thankfully, it did not appear as if there was strong racial bias within this dataset, 'B' variable

	or percent African Americans correlation = .33 (weak positive correlation). • Multicollinearity (two explanatory variables in regression model that are highly related) should be avoided because they result in poor/unstable estimates. IE - don't pick two predictor variables that correlate strongly with each other. (https://stats.stackexchange.com/questions/1149/is-there-an-intuitive-explanation-why-multicollinearity-is-a-problem-in-linear-r/1150#1150)
Thoughts on Future Project	• Last semester (in Data601), I analyzed three datasets, and my final project looked at correlations between variables in predicting poverty: https://github.com/Colsai/Education_DS_Projects/blob/main/New_York_HS_Pov.ipynb • At present, my hope is to look at another education dataset, to try and make predictions on school performance based on predictive variables.