Instructions: Please read the following instructions thoroughly

- For the entire assignment, use `Python` for your analysis. Write your code in a Jupyter Notebook named as `[your-student-ID]_hw2.ipynb` (e.g., `2023-20000_hw2.ipynb`). The use of `R` is not allowed. You are allowed to use any libraries in `Python`. Type up your report and save as PDF named as `[your-student-ID]_hw2.pdf`. We do not allow the submission of a photo or a scanned copy of hand-written reports.

- Please upload the two files, your report in PDF and the code in Jupyter Notebook on eTL **without zipping**. Submissions via email are not allowed. The violation of the filename or submission instruction will result in the penalty of 5 points. No worries if there are additional numbers at the end of the filename that appear when you submit more than once.

- You can discuss the assignment with your classmates but each student must write up his or her own solution and write their own code. Explicitly mention your classmate(s) you discussed with or reference you used (e.g., website, Github repo) if there is any. If we detect a copied code without reference, it will be treated as a serious violation of the student code of conduct.

- We will apply a grace period of late submissions with a delay of each hour increment being discounted by 5% after the deadline (i.e., 1-minute to 1-hour delay: 95% of the graded score, 1 to 2-hour delay: 90% of the graded score, 2 to 3-hour delay: 85%, so on). Hence, if you submit after 20 hours post-deadline, you will receive 0 points. No excuses for this policy, so please make sure to submit in time.

1. We will predict the number of applications received using the other variables in the `College` data set.

   (a) [10 pts] Randomly split the data set into a training set (90%) and a test set (10%). Fit a linear model using least squares on the training set, and report the test error obtained.

   (b) [20 pts] Fit a ridge regression model on the training set, with λ chosen by 10-fold cross-validation using the training set. Report the test error obtained.

   (c) [20 pts] Fit a lasso model on the training set, with λ chosen by 10-fold cross-validation using the training set. Report the test error obtained, along with the number of non-zero coefficient estimates.

   (d) [10 pts] Comment on the results obtained. How accurately can we predict the number of college applications received? Are the models resulting from these three approaches different from each other?

2. We will now try to predict per capita crime rate in the `Boston` data set.

   (a) [20 pts] Try the forward step-wise selection, the backward step-wise selection, the lasso, and ridge regressions. Use cross-validation if needed for choosing a tuning parameter. Present and discuss results for the approaches that you consider.

   (b) [10 pts] Propose a model (or set of models) that seem to perform best on this data set, and justify your answer. Make sure that you are evaluating model performance using cross-validation, as opposed to using training error.

   (c) [10 pts] Does your chosen model involve all of the features in the data set? Why or why not?