3. (*15 pt*) In this question, you will analyze a dataset of your choice. You have two options here:

   (I) Choose a real dataset with a quantitative response y and $p \geq 30$ predictors.

   I recommended that you search for a dataset on UCI Machine Learning Repository. These tend to have clean data that is ready for analysis. Data from other sources such as Kaggle, Havard Dataverse and Google Dataset Search are also acceptable. These may require you to do some cleaning, which digresses from the questions.

   (II) Generate your own data with a qualitative response y and $p \geq 30$ predictors. Ensure that at least two of the p predictors are linearly dependent. (If you choose option (I), you don't have to verify the presence of any linear dependencies.)

   Choose wisely and answer the following questions.

   (a) (*2 pt*) Did you choose option (I) or (II)?

   If you chose option (I), describe your data and the source. What is your response variable, and what are the predictors? You need not enumerate all predictors, but write a few lines giving context for the dataset.

   If you chose option (II), give an overview of how you generated the data and how you introduced dependencies between predictors.

   (b) (*2 pt*) Split your data into training and test (use either a 80/20 or 90/10 split i.e., 80% points in training and 20% points in test, or 90% in training and 10% points in test). Ensure that the number of data points in your training subset is greater than the number of predictors i.e., $n_{\text{train}} > p$.

   With the `regsubsets` function in the `leaps` package, use forward selection or backward selection (choose either) to fit a reduced linear model (reduced as in, your final model uses fewer than p predictors) on the training data. Report the number of predictors used in your linear model and the error on the test data. Note that you do not use the test data during forward/backward selection.

   (c) (*3 pt*) Using the subset of predictors identified in part (b), fit a ridge regression model on the training data with a range of values for the regularization parameter λ. You can use the `grid` function (from page 275 of ISLR) to get a range of λ values. Use k-fold cross validation with $k = 10$ folds to choose the optimal λ, but feel free to decrease the number of folds if your data is too large. Note that you will have to standardize the features before performing ridge regression.

   Generate and submit a plot that resembles the left panel Figure 6.4 in ISLR i.e., λ on the horizontal axis and the estimated (standardized) coefficients on the vertical axis (a different line for each predictor). No need to generate the right panel.

   Note that you did not use the test data at all for this sub-problem.

   (d) (*2 pt*) Find the value of λ that provides the smallest cross-validation error in part (c). Using this λ, find the test error of this ridge regression model. Compare this with the error in part (b).

   (e) (*3+2+1 pt*) Repeat parts (c-d) using lasso regression. What features, according to lasso regression, have a non-zero estimated coefficient?

***