Question 1

Use the dataset donation (*donation.csv*) where "Donation" is the outcome variable and the predictors are listed in the table below. Use R to analyze the data:

1. Create a summary statistics table.
2. Print the frequency table for the factor variable Homeowner. Compare variable averages for different values of Homeowner. Do you find any interesting patterns?
3. Partition the data into 70% training and 30% validation (test) using seed 310.
4. Using the training set, run a linear regression model with all the variables.
5. What are the significant predictors? How do you interpret them?
6. Predict the outcome for the training and test sets.
7. Calculate the residuals for train and test. Are the errors heteroskedastic or homoscedastic?
8. What is RMSE for test and train? Do you think there is overfitting problem?
9. Run a best subset model. Which model is the best model?
10. Do the model selection using backward selection. Which model has the highest adjusted R^2? For a model with 6 predictors, what are the selected 6?
11. Use Ridge regression and find lambda.min? What is MSE for lambda.min?
12. Use Lasso regression and find lambda.min? What are the coefficients for lambda.min?
13. In your own word explain what is the difference between lambda.min and lambda.1se? How do we choose one over the other?
14. Which model do you think is better?
15. If you are consulting a fund-raising campaign, what strategies do you suggest based on the insights from the analyses?

Predictors

Homeowner	**Categorical** (Y = Yes; N = No)
NUMCHLD	Number of children
INCOME	Household income
GENDER	**Categorical** (M = Male; F = Female)
WEALTH	Wealth Rating
HV	Average Home Value in potential donor's
ICmed	Median Family Income in potential donor's
ICavg	Average Family Income in potential donor's
IC15	Percent earning less than 15K in potential
NUMPROM	Lifetime number of promotions received to date
RAMNTALL	Dollar amount of lifetime gifts to date
MAXRAMNT	Dollar amount of largest gift to date
LASTGIFT	Dollar amount of most recent gift
TOTALMONTHS	Number of months from last donation to July 1998
TIMELAG	Number of months between first and second gift
AVGGIFT	Average dollar amount of gifts to date

Output Variable

Donation	Amount donated (in $)

Question 2

Consider the ebay dataset (*eBay.csv*) where "Competitive" is the outcome variable which indicates whether an auction is competitive (coded as 1) or not (0). Use logistic regression for this classification problem.

1. Plot sellerRating vs. OpenPrice with different colors for Competitive with a single linear smoothing line. Create labels and titles for the plot.
2. Using summaryBy() function from doBy package, compare the average of OpenPrice and Duration for Competitive and not Competitive. What do you think about the finding?
3. Split the data into 70% training and 30% test using seed 310.
4. Run the logistic regression on the training set.
5. How many predictors are significant? Interpret the marginal effect of OpenPrice?
6. Predict the training and test sets and get the confusion matrices.
7. Calculate the error rates.
8. Do we have the overfitting problem?
9. Create the ROC curve.
10. What can you say about the ROC curve?
11. Suppose you work for eBay. Based on the information you found in this problem, what do you think about the duration of the auctions?