# Assignment 5 - Classification

Answer Key

"When life gives you lemons, don't make lemonade. Make life take the lemons back! Get mad!" - Cave Johnson

For this assignment, you'll be using the lemons dataset, which is a subset of the dataset used for a Kaggle competition described here: https://www.kaggle.com/c/DontGetKicked/data. Your job is to predict which cars are most likely to be lemons - or, in other words, "bad buys." You'll want to read in the lemons dataset from the course website. As always, please put any sentences for your question answers OUTSIDE of code chunks.

Read in the data and examine it using the summary function.

1. Calculate the proportion of lemons in the training dataset using the `IsBadBuy` variable. What proportion of cars in this data are lemons?
2. Calculate the proportion of lemons by Make. Which make of car has the highest proportion of lemons and what is that proportion?
3. Examine other independent variables' potential relationship with your dependent variable. For each potential IV, include a sentence stating whether or not it appears to have a potential relationship with your DV (and how you know that). The goal for this problem is to arrive at ONE additional IVs (beyond Make) to include in your logistic model below.
4. Before running a logistic regression, you need to make your outcome variable into a factor variable and split the data into training and testing. Use set.seed(5634) and split your data to create testing/training datasets.
5. Now, predict the probability of being a lemon using a logistic regression, using Make and at least one other covariate/independent variable of your choosing (Note: this should be informed by your examinations from #3 above). In this chunk you should set the model, set up the recipe, define the model, put the workflow together, fit the results, and then present your tidy table (these are all the steps we went through together in the code - nothing new here).
6. Choose two variables from your multiple logistic regression model above that were statistically significant predictors (at alpha=.05) of a car being a bad buy and provide one sentence for each interpreting the regression estimate (remember - you can convert the estimate to an odds ration by exponentiating it; be sure to account for the scientific notation of your estimate before exponentiating).
7. Use your logit model results to make predictions in your testing dataset and create a confusion matrix. How many "NOs" did you predict correctly? How many "YESs" did you predict correctly?
8. Calculate the accuracy, sensitivity and specificity of your model. Provide one sentence interpreting each of these three values.
9. Calculate the AUC for the predictions from the ROC based on the logit model. Provide one sentence interpreting this roc_auc.