# Homework 7

For this homework you will create a python notebook (`.ipynb` file). This file should then be uploaded to wolfware in the assignment link! You can use colab or our jupyter hub for this assignment as we aren't using pyspark here.

## Goal

The purpose of this homework is to practice fitting MLR and logistic regression models (including penalized or regularized models). Most homework assignments will have a part that pushes you beyond what was in the lectures! Learning to search for the right questions and browsing stackoverflow are really life skills that we should hone :) **Be sure to include markdown text describing what you are doing, even when not explicitly asked for!**

## Data

We will use a dataset from the UCI Machine Learning Repository. This data set is about wine quality (we played around with this data in class at some point I think). You can learn more about the data here.

The data description describes the following variables:

Input variables (based on physicochemical tests)

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

- Rather than try to predict `quality`, let's make our target variable for fitting multiple linear regression type models `alcohol`.
- For fitting logistic regression type models we'll use the type of wine as the response variable.

## To Do:

Create a document that goes through your process of reading the data, combining it, manipulating/creating any variables, and fitting and choosing a final model for both the multiple linear regression modeling and the logistic regression modeling described below.

### Read in and Combine Data

- Read in the `winequality-red.csv` and `winequality-white.csv` files available on the UCI machine learning repository site.
- Combine these two datasets and create a new variable that represents the type of wine (red or white)

### Split the Data

- Split up the data set into a training and test set. For this, I want you to use stratified sampling to make sure that you have a similar proportion of white and red wines in the training and test sets. This can be done with the `train_test_split()` function (see the help).

### Regression Task (`alcohol` as Response)

### Train Models

- Fit four different multiple linear regression models.
  - At least one should include interaction terms
  - At least one should include some polynomial terms (you may want to standardize your predictors but that is up to you)
  - Use CV to select your best MLR model
- Fit a LASSO model with a set of predictors of your choosing
  - Use at least five predictors
  - Use CV to select the tuning parameter
- Fit a Ridge Regression model with a set of predictors of your choosing
  - Use at least five predictors
  - Use CV to select the tuning parameter
- Fit an Elastic Net model with a set of predictors of your choosing
  - Use at least five predictors
  - Use CV to select the tuning parameters

### Test Models

- Using your four selected models, compare their performance on the test set.
  - Do so using RMSE as your model metric
  - Do so using MAE as your model metric

### Classification Task (Wine Type as Response)

- Repeat the training and testing done previously but use logistic regression models.
- Use log-loss or negative log-loss as your metric for choosing models during the training process
- During the testing portion, compare your models on both log-loss and accuracy

## Files to Submit

- Make sure all cells are run
- Click on File –> Save and Export Notebook As... –> HTML

That's it! Upload both the `.ipynb` and `.html` files to the assignment link.