Project

For this project you will create a pdf or HTML file with text, code, and results (perhaps using an R Markdown output, a Jupyter Notebook, or a .qmd file output). This output file (pdf or HTML) and the corresponding file used to create the output (.Rmd, .ipynb, or .qmd) should then be uploaded to wolfware in the assignment link!

# Goal

The purpose of this project is to create a report where you read in data, fit models using concepts from the class, and choose between several different models.

## Data (**Must be Approved by 3/26!**)

* Find a data set you can fit supervised learning models with
  + You cannot use the datasets we’ve been using in class
  + You can’t use a super simple data set such as the iris data set nor data on wine quality.
  + There are many places with free data out there such as the UCI machine learning repository and kaggle
* You can focus on a classification task or a regression task, I’ll leave that to you!
* Please read your data in via a URL or include the data as a file in your submission.
* **Please send me an email with a link to the data ASAP and I will let you know if the data is OK to use for the project.**

## To Do:

Create a document that goes through your process of reading the data, any basic data cleaning/transformations, splitting the data, and fitting and choosing a final model. Details are given below.

### Read in the Data

* Give a brief introduction to the data and the source of the data.
* State your goal for the project (modeling something). Be specific on why you want to model the variable.
* Read the data in via a URL (or locally but then you must submit the data with your submission)

### Data Cleaning & Transformations

* Go through a brief process where you make sure that each variable is read in correctly (correct data type), check for missing values, check for valid data values (perhaps by summarizing each column and seeing if you have reasonable values), removing observations where needed (give an explanation as to why), and doing any data transformations you deem necessary.

### Split the Data into a Train and Test Set

* Use whatever reasonable proportion you’d like.

### Training Models

In the course so far, some of the models we’ve looked at are

* kNN
* (Regularized) Linear Regression & Logistic Regression Models
* GAMs (using piece-wise polynomial regression, local regression, or splines)
* Single tree models
* Ensemble tree models
* Support vector machines

You’ll fit one model from each of these model types (I’ll let you determine the details).

For **each model type**, do the following:

1. Describe the model. For instance,
   * is the model a parameteric model? non-parametric?
   * does it have tuning parameters? if so, what are they?
   * can we use the model for inference?
   * does the model perform variable selection?
   * do we need to standarize the predictors?
2. Fit the models on the training set, selecting tuning parameters where needed using a reasonable metric.
3. Select the ‘best’ model and refit it to the training data.

### Testing Models

Test each of your models on the test set. Report appropriate metrics and discuss which model you prefer and why (perhaps simplicity, perhaps predictive ability, etc.)

Fit your overall best model on the entire data set.

Interpret this model as best you can. That is, if you can do inference with the model, do so. If you can simply report variable importance measures, do that. If you can produce plots that describe the relationships, do that!

That’s it! Submit the appropriate files.