STAT 8330: **Project 1**

Due on October 26, 2020

Instructions: This project will consist of competing in a Kaggle competition – House Prices: Advanced Regression Techniques. The details of the competition can be found at `https://www.kaggle.com/c/house-prices-advanced-regression-techniques`. The specific requirements for this class project are detailed below.

Data set: The data found at `https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data` consists of information on 79 explanatory variables (predictors) describing (almost) every aspect of residential homes in Ames, Iowa, that will be used to predict the final price of each home. As described at `https://www.kaggle.com/c/house-prices-advanced-regression-techniques#evaluation`, the metric here is root mean squared error (RMSE).

Logistics:

- You will work individually on this project.
- You will create a Kaggle account. Although you can submit multiple entries to the competition (see the competition rules), you are allowed to submit only 5 models total to the Kaggle competition for this class!! (Obviously, I know you can cheat this, but really, what is the point? For this project, I am more interested in WHY you chose your models than exactly how well you do!) You MUST provide results from your best 3 models from the Kaggle evaluation (e.g., screen shots).

  Because I am limiting you to 5 models, you must justify in your report how you selected these models in terms of RMSE **based on the training data alone**. It is **NOT** acceptable to use the Kaggle results as a way to train your models – so, your final model has to be justified solely based on the training data!!!

  Lastly, note that you will have to submit your predictions to the Kaggle site according to their special format described under the Evaluation tab on the competition website given above. (Note, you can find various tutorials as well – see some of the other "getting started" competitions, such as the Titanic survial one.)

Tasks:You will turn in a project report that contains the following:

- a very brief introduction to the problem (no more than 1 paragraph).
- a description of any basic data analysis/description/plots that you feel are important; these must be relevant to your data analysis or don't include them! (If you do something that is helpful, it will help your grade.) Examples include, feature engineering, plots, etc. (good projects will have something to report here)
- a brief description of how you decided on your three models and any pre-processing you did to the data.
- a brief description of your models and your results in terms of RMSE (include screen shots from your submission to the competetion!)
- the exact R commands necessary to obtain the exact results from your model in terms of the training data set. Include any preprocessing commands (I need to be able to duplicate your results exactly!). This will be a separate file uploaded to the Canvas site.
- a very brief conclusion.