HW 2, STAT 452
Due: Thursday, February 18

Reading: Chapter 5, pp. 175–183, from *An Introduction to Statistical Learning*
Chapter 2 from *Hands-on Machine Learning*

Directions: Please submit your completed assignment to Blackboard. The assignment should be completed using R Markdown and rendered to an HTML or PDF format. Note that Blackboard will not accept HTML files. One workaround is to first zip your HTML file, and then submit the zipped file to Blackboard.

```
library(tidyverse) # load tidyverse packages (ggplot2, dplyr, ...)
library(AmesHousing) # load Ames housing data set
ames <- make_ames() # set up data frame
```

Exercise 1. Answer the following as True or False:

(a) The mean squared error (MSE) is an appropriate and commonly used measure of performance for classification tasks, such as predicting whether or not a person has a disease.

(b) The root mean squared error (RMSE) is in the same units as the response data. So if the response variable is price in US dollars, then the RMSE can also be interpreted in terms of US dollars.

(c) The best measure of predictive performance for a multiple linear regression model is the coefficient of determination, R^2, which we can easily look up when we run the `summary()` function on a linear model object.

Exercise 2

(a) Split that `ames` data frame into a 70% training and 30% test set.

(b) Fit the following three regression models on the training set:

```
Sale_Price ~ Gr_Liv_Area + Year_Built

Sale_Price ~ Gr_Liv_Area + Year_Built + TotRms_AbvGrd

Sale_Price ~ Gr_Liv_Area + Year_Built + TotRms_AbvGrd + Overall_Cond
```

(c) Make predictions on the test set and compute the RMSE for the three regression models. According to the RMSE, which model has the best predictive performance? Provide an interpretation of the RMSE for this model.

(d) Compute the mean absolute error (MAE) for the three regression models. How do the MAE results compare with the RMSE?

(e) For the regression model with the best predictive performance, make a plot of the predicted versus actual values on the test set.