HW 3, STAT 452
Due: Thursday, February 25

Reading: Chapter 2, pp. 29–36, and Chapter 5, pp. 181–183, from *ISLR*
Chapter 4 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.

Bonus! Complete this assignment using RStudio Cloud and earn 5 points extra credit. See announcement on Blackboard for more info.

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

Exercise 1. Answer the following as True or False. If False give a brief explanation.

(a) A primary goal when fitting a statistical model is to minimize the MSE on the training data.

(b) Suppose one fits a simple linear regression model to a scatterplot that shows a nonlinear, possibly quadratic association between x and y. Then this would be an example of underfitting.

(c) Suppose one fits a multiple linear regression model with $p = 50$ predictor variables to a training set with $n = 100$ observations. A test set with $n = 30$ points is withheld. The RMSE on the test set is much larger than the RMSE on the training set. Then it would be reasonable to conclude that that the regression model with the $p = 50$ predictors is overfitting the data, and that some predictor variables should be removed from the model to improve prediction performance.

(d) Suppose one fits a multiple linear regression model with $p = 50$ predictor variables to a data set with $n = 100$ observations. The R^2 for the model, given from the regression summary, is close to 1. Therefore, it's reasonable to conclude that the model is a good fit to the data, and will do a remarkable job at making predictions for future values of the response variable.

Exercise 2. Use the `train()` function from the `caret` package to perform 10-fold cross-validation for the following three linear regression models:

```
Sale_Price ~ TotRms_AbvGrd
```

```
Sale_Price ~ TotRms_AbvGrd + Year_Built
```

```
Sale_Price ~ TotRms_AbvGrd + Year_Built + Bldg_Type
```

Make sure to set the same random seed with `set.seed()` before you run `train()` for each model. Which model has the best predictive performance? Explain.