HW 8, STAT 452

Due: Friday, May 7

Reading: Chapter 8 from *An Introduction to Statistical 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.

```r
# load packages
library(tidyverse)
library(rpart)
library(randomForest)
library(vip)
```

Exercise 1. Consider the Abalone Data Set, which can be accessed by loading the following package:

```r
library(AppliedPredictiveModeling)
data(abalone)
```

The data consist of measurements of the type (male, female and infant), the longest shell measurement, the diameter, height and several weights (whole, shucked, viscera and shell) from 4177 abalones. The goal, or machine learning task, is to predict the number of rings of the abalone from these attributes.

The age of an abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task. Thus, it would be useful to use other measurements (length, diameter, weight), which are easier to obtain, to predict the age.

(a) It is always good to start by exploring the data a little. Use the `summary()` function to compute summary statistics for the variables. Make some scatterplots between the response variable, `Rings`, and some of the other predictor variables. Then write a few sentences describing the relationships. (You don't need to spend that much time on this question. Just make some graphs and compute summary statistics to help you get an understanding of the data.)

(b) Randomly split the `abalone` data frame into a 70% training and 30% test set. Make sure to use `set.seed()` so that your results are reproducible.

(c) Use `lm()` to fit a multiple linear regression model **on the training set** with `Rings` as the response, and all the other variables in the data frame as predictors. Use `summary()` to print the regression output (coefficient table).

(d) Use `rpart()` to fit a regression tree **on the training set** with `Rings` as the response, and all the other variables in the data frame as predictors. Make a plot of the regression tree.

(e) Use `randomForest()` to fit a random forest model **on the training set** with `Rings` as the response, and all the other variables in the data frame as predictors. Make a variable importance plot.

(f) Make predictions on the test set and compute the RMSE for the three models (multiple linear regression, regression tree, and random forests). According to the RMSE, which model has the best predictive performance? Which model has the worst predictive performance? Write a few sentences summarizing and interpreting the cross-validation results.

```r
# function to compute RMSE
RMSE <- function(y, y_hat) {
  sqrt(mean((y - y_hat)^2))
}
```

(g) Make plots of the predicted versus actual values on the test set for each of the three models; add the 1-1 reference line to each plot. Comment on why the patterns in the plot of the predicted versus actual values for the regression tree model look different than the random forest and linear regression models?