HW 7, STAT 452
Due: Thursday, April 15

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.

```
# load libraries
library(tidyverse)
library(rpart)
```

Exercise 1. For this exercise, use the 2016 election data for US counties (see lecture 10 slides for a description of the data set and variables):

```
county_votes <- readRDS(url("https://ericwfox.github.io/data/county_votes16.rds"))
```

(a) Run the following line of code to create a new variable (column) in the data frame called `result`, which is a factor with two levels (`trump`, `clinton`) that indicate which candidate won the county. Essentially, this is a recoding of the binary (1/0) response variable `trump_win` with more descriptive names for the categories.

```
county_votes$result <- factor(county_votes$trump_win,
                              levels = c(1,0), labels=c("trump", "clinton"))
```

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

(c) Use `rpart()` to fit the following classification tree on the training set:

```
result ~ obama_pctvotes + pct_pop65 + pct_black + pct_white + pct_hispanic
            + pct_asian + highschool + bachelors + income
```

(d) Make a plot the classification tree estimated in part (c). Based on this plot, which predictor variable appears to be the most important, or useful for predicting the election result in each county?

(e) Make predictions for the response `result` on the test set, and compute the confusion matrix. Then use the confusion matrix to compute the accuracy (percent correctly classified), percent of Trump wins correctly classified, and percent of Clinton wins correctly classified.