Extra Credit Assignment, STAT 452
Due: Friday, May 7

Instructions: This extra credit assignment is worth 5 points. Please submit your completed assignment to Blackboard. The assignment should be completed using R Markdown and rendered to an HTML or PDF format.

```
# load packages
library(tidyverse)
library(randomForest)
library(ranger)
```

For this assignment, we'll use a subset of data from the MNIST data set (lecture 18). Run the following command to load the training and test set:

```
mnist_train_49 <- readRDS(url("https://ericwfox.github.io/data/mnist_train_49.rds"))
mnist_test_49 <- readRDS(url("https://ericwfox.github.io/data/mnist_test_49.rds"))
```

The objective here is to train a random forest classifier that can predict whether a handwritten digit is either a 4 or a 9.

The data frame `mnist_train_49` contains 5000 rows and 785 columns. Each row represents an image of a handwritten 4 or 9. The first column `y` is the response variable with the actual labels:

```
dim(mnist_train_49)

## [1] 5000  785

table(mnist_train_49$y)

##
##    4    9
## 2467 2533
```

The other 784 columns (features) represent the gray-scale intensities of each pixel in the 28x28 image (a number ranging from 0 to 255). The data frame `mnist_test_49` has similar structure and will be used as the test set.

```
dim(mnist_test_49)

## [1] 1000  785

table(mnist_test_49$y)

##
##   4   9
## 490 510
```

(a) Run the following code to look at the images of the first two digits in the training set.

```
m1 <- mnist_train_49[1, ]
m1 <- as.numeric(m1[-1])
image(matrix(m1, 28, 28)[, 28:1],
      col = gray.colors(12, rev = TRUE), xaxt="n", yaxt="n")
```

```
m2 <- mnist_train_49[2, ]
m2 <- as.numeric(m2[-1])
image(matrix(m2, 28, 28)[, 28:1],
      col = gray.colors(12, rev = TRUE), xaxt="n", yaxt="n")
```

(b) Fit a random forest classifier that predicts the digit labels (4's and 9's) using the 784 pixel intensities from an image as features (predictor variables). Use the training set (`mnist_train_49`) to estimate the model.

(c) Make predictions for the digits on the test set (`mnist_test_49`). Compute the confusion matrix. What is the overall accuracy (percent correctly classified)? What percent of 4's were correctly classified? What percent of 9's were correctly classified?