HW 5, STAT 452
Due: Thursday, March 18

Reading: Chapter 2, pp. 39–42, and Chapter 4, pp. 163–167, from *An Introduction to Statistical Learning*

```
library(palmerpenguins)
library(tidyverse)
library(class) # library with knn function
```

Exercise 1. Explain why predictor variables are usually standardized prior to running the kNN algorithm. Describe two ways that numeric variables can be standardized.

Exercise 2. Refer to `lecture11_code.Rmd` when completing this exercise.

(a) Use `ggplot2` to make a scatter plot with `bill_lengh_mm` on the x-axis and `body_mass_g` on the y-axis. Color the points according the `species`.

(b) Run the following code to prepare the data for the kNN classifier. The code standardizes the two predictors `bill_lengh_mm` and `body_mass_g` between 0 and 1, and also removes any missing data entries.

```
# function to standardize numeric predictors between 0 and 1
standardize <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}
```

```
penguins2 <- penguins %>%
  select(species, bill_length_mm, body_mass_g) %>%
  na.omit() %>%
  mutate(bill_length = standardize(bill_length_mm)) %>%
  mutate(body_mass = standardize(body_mass_g)) %>%
  select(-bill_length_mm, -body_mass_g)
```

(c) Split the data into a 70% training and 30% test set. Make sure to set a random seed so that your results are reproducible.

(d) Use the `knn()` function to run the kNN algorithm using $k = 1$ (the nearest neighbor). Make a confusion matrix between the actual and predicted species classes on the test set, and compute the overall accuracy (percent correctly classified).

(e) Use the `knn()` function to run the kNN algorithm using $k = 5$ (the 5 nearest neighbors). Make a confusion matrix between the actual and predicted species classes on the test set, and compute the overall accuracy (percent correctly classified).

(f) How does accuracy (percent correctly classified) on the test set compare when $k = 1$ and $k = 5$?