STA130: Week 10 R Demo

Palmer Penguins

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(palmerpenguins)
```

Goal: We want to try to predict the species of a penguin, based on the information we know about them

```
library(rpart)
library(partykit)
```

```
## Loading required package: grid
```

```
## Loading required package: libcoin
```

```
## Loading required package: mvtnorm
```

```
# First, we'll restrict attention to observations that don't have any missing values
summary(penguins) ## New function to give a quick overview of each of the variables in a tibble
```

```
##       species          island    bill_length_mm  bill_depth_mm  
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10  
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60  
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30  
##                                  Mean   :43.92   Mean   :17.15  
##                                  3rd Qu.:48.50   3rd Qu.:18.70  
##                                  Max.   :59.60   Max.   :21.50  
##                                  NA's   :2       NA's   :2      
##  flipper_length_mm  body_mass_g       sex           year     
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007  
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007  
##  Median :197.0     Median :4050   NA's  : 11   Median :2008  
##  Mean   :200.9     Mean   :4202                Mean   :2008  
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009  
##  Max.   :231.0     Max.   :6300                Max.   :2009  
##  NA's   :2         NA's   :2                                 
```

```
# Now, we'll divide our data into training/testing datasets
# Set up
```

```
set.seed(17);


# Create training and testing datasets

# How many observations are there in each of the training and testing datasets?
```

```
# Let's build a tree using only geographic information to predict penguin species

# What is the difference between type="simple" and type="extended" for visualizing a classification tre

# How can we visualize what is going on behind the scenes?


# Let's build a second tree using only physiological information to predict penguins species

# Were all of the candidate predictors used to make splits in tree2?


# Now let's build a third tree which allows for all variables (apart from species) to be used to predic

# What's weird/wrong with the tree above?


# Let's try that again
```

Now let's compare our three trees!

```
# Make predictions for test observations based on tree1

# What is the accuracy for tree1 based on testing data?

# Can we calculate the sensitivity/specificity for this tree?

# Which type of penguins are hardest to classify based on this tree?


# Make predictions for test observations based on tree2


# What is the accuracy for tree1 based on testing data?



# Make predictions for test observations based on tree3

# What is the accuracy for tree1 based on testing data?

# What do you notice about the confusion matrices for trees 2 and 3?


# Which tree would you prefer to use: tree1 or tree2/3?
```