

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
```

```
penguins_clean <- penguins %>% filter(!is.na(bill_length_mm) & !is.na(bill_depth_mm) & !is.na(flipper_length_mm))
summary(penguins_clean)
```

```
##      species      island  bill_length_mm  bill_depth_mm
##  Adelie   :151  Biscoe   :167  Min.      :32.10  Min.      :13.10
##  Chinstrap: 68  Dream    :124  1st Qu.:39.23  1st Qu.:15.60
##  Gentoo   :123  Torgersen: 51  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
## 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  : 9   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
```

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

```
# Set up
```

```
set.seed(17);
```

```
n <- nrow(penguins_clean)
```

```
training_indices <- sample(1:n, size=round(0.8*n))
```

```
penguins_clean <- penguins_clean %>% rowid_to_column() # adds a new ID column
```

```
# Create training and testing datasets
```

```
train <- penguins_clean %>% filter(rowid %in% training_indices)
```

```
test <- penguins_clean %>% filter(!rowid %in% training_indices)
```

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

```
nrow(train)
```

```
## [1] 274
```

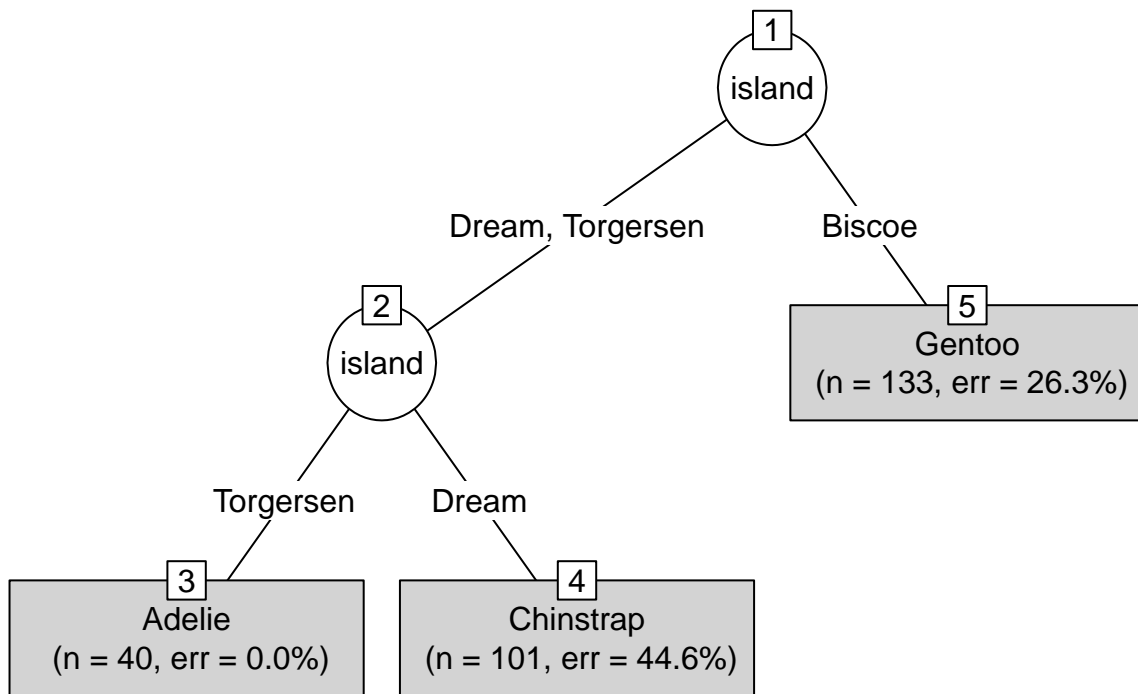
```
nrow(test)
```

```
## [1] 68
```

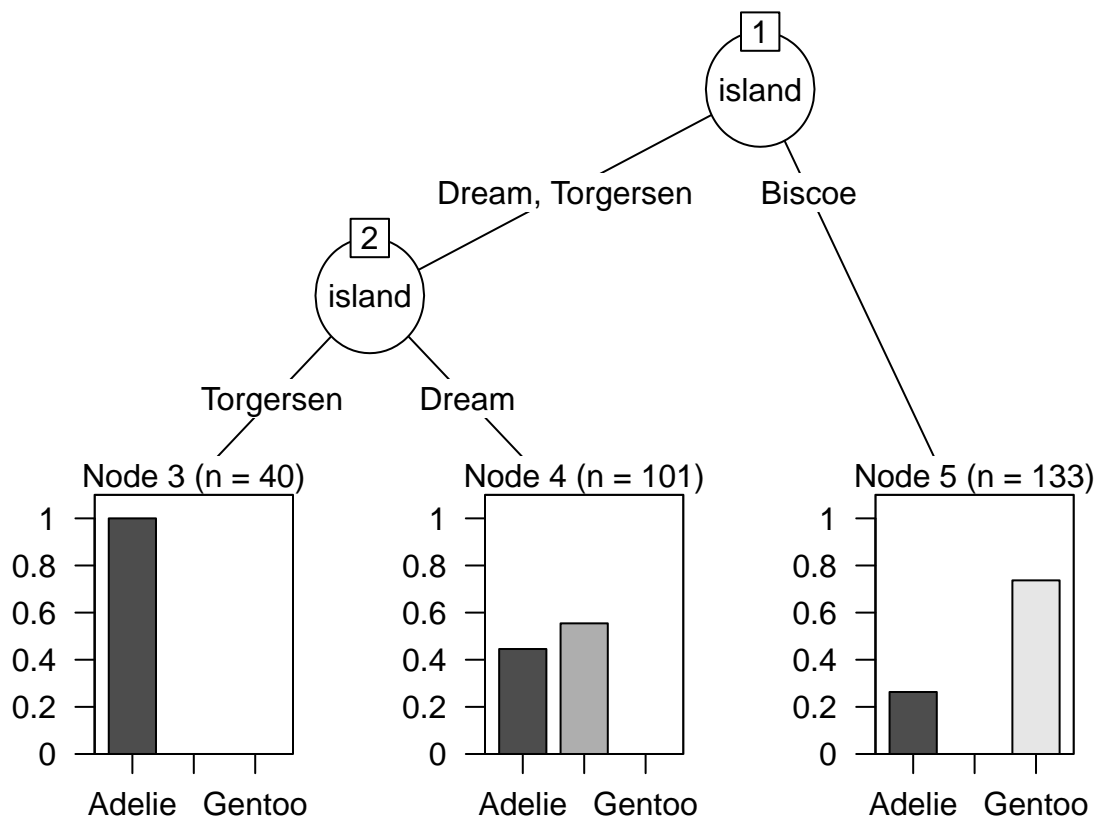
```
# Let's build a tree using only geographic information to predict penguin species
```

```
tree1 <- rpart(species ~ island, data=train)
```

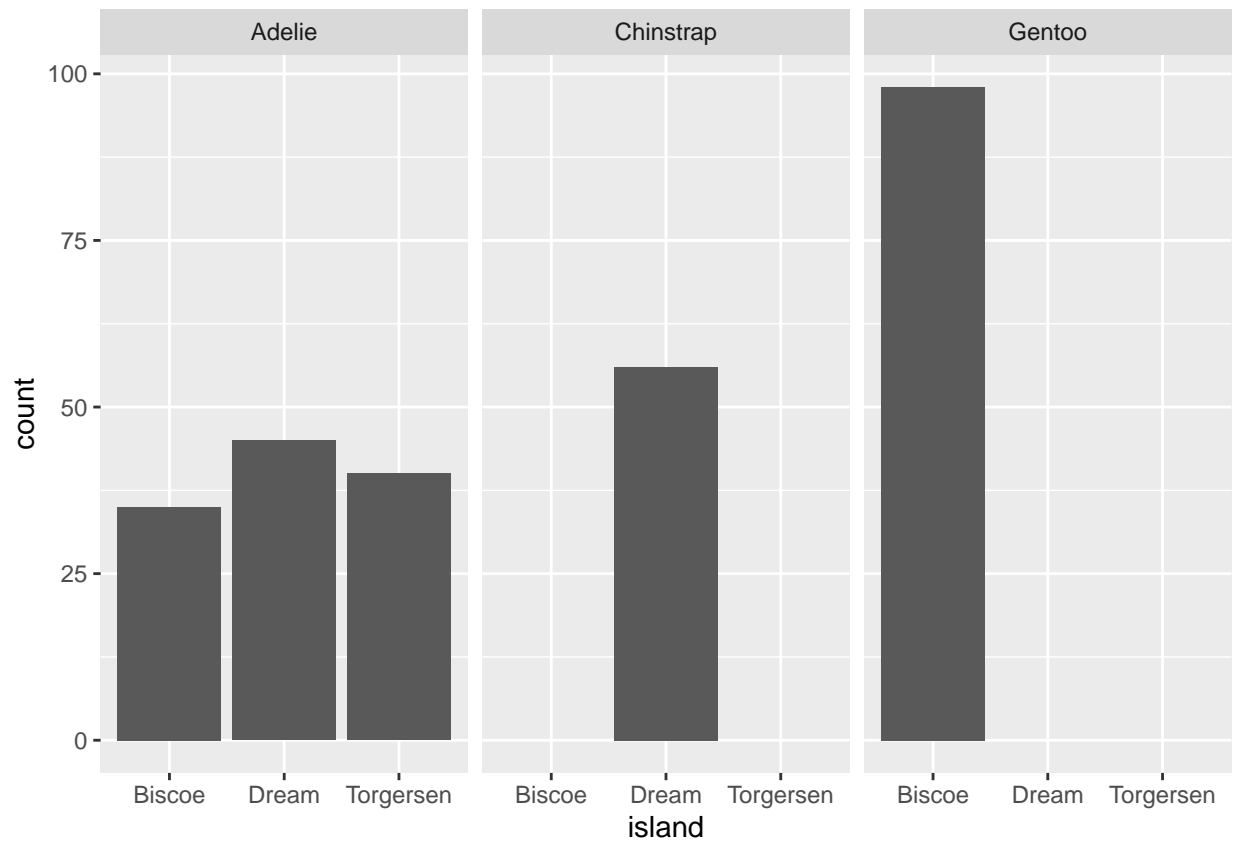
```
plot(as.party(tree1), type="simple")
```



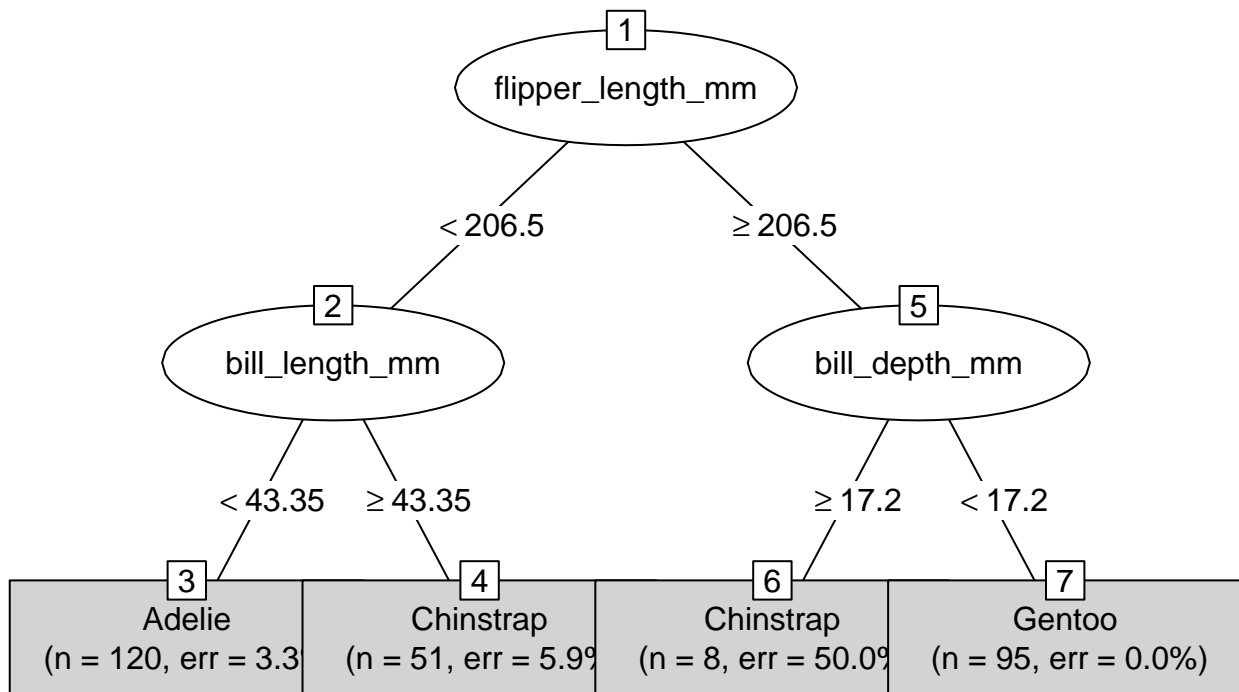
What is the difference between `type="simple"` and `type="extended"` for visualizing a classification tree?
`plot(as.party(tree1), type="extended")`



```
# How can we visualize what is going on behind the scenes?
train %>% ggplot(aes(x=island)) +
  geom_bar() + #coord_flip() +
  facet_wrap(~species)
```

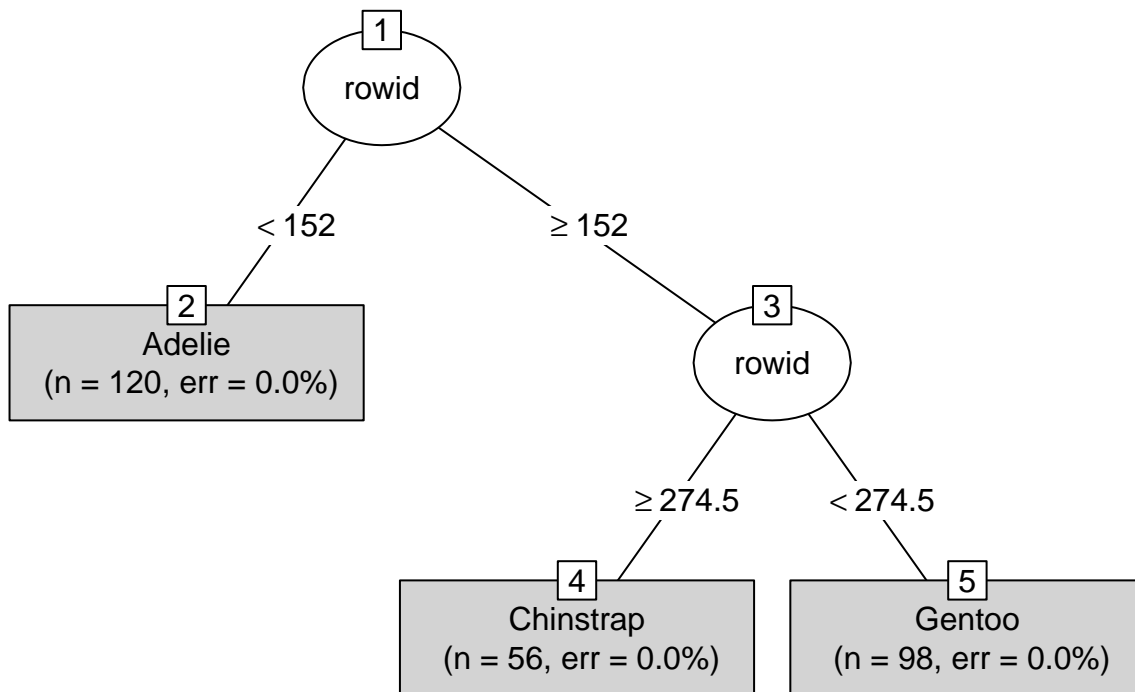


```
# Let's build a second tree using only physiological information to predict penguins species
tree2 <- rpart(species ~ bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g + sex, data=t)
plot(as.party(tree2), type="simple")
```



```
# Were all of the candidate predictors used to make splits in tree2?
#No! It turns out that only some of the candidate predictors led to splits which improved the purity o

# Now let's build a third tree which allows for all variables (apart from species) to be used to predic
tree3 <- rpart(species ~ ., data=train)
plot(as.party(tree3), type="simple")
```



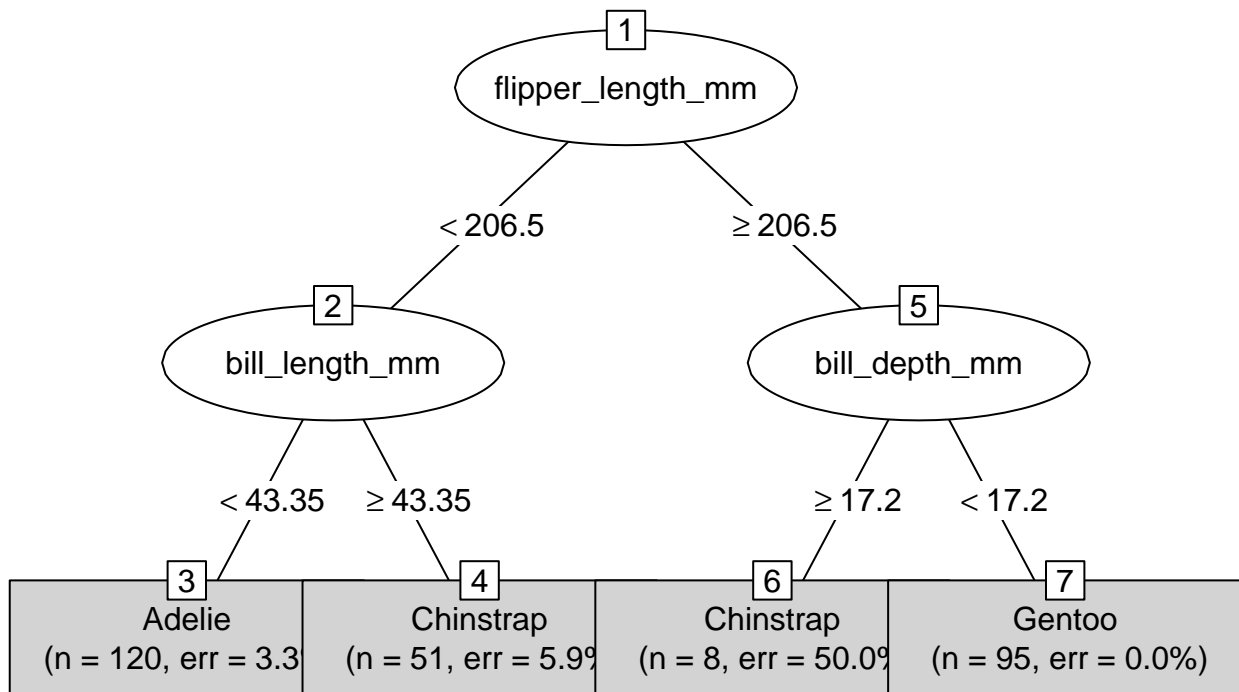
What's weird/wrong with the tree above?

Let's try that again

```

tree3 <- rpart(species ~ ., data=train %>% select(-rowid))
plot(as.party(tree3), type="simple")

```



Now let's compare our three trees!

```
# Make predictions for test observations based on tree1
test_preds_1 <- predict(tree1, newdata=test, type="class")
head(test_preds_1)
```

```
##      1      2      3      4      5      6
## Adelie Adelie Adelie Adelie Adelie Gentoo
## Levels: Adelie Chinstrap Gentoo
```

```
m1.test <- table(test_preds_1, test$species)
m1.test
```

```
##
## test_preds_1 Adelie Chinstrap Gentoo
##      Adelie      11         0         0
##      Chinstrap    11        12         0
##      Gentoo       9         0        25
```

```
# What is the accuracy for tree1 based on testing data?
sum(diag(m1.test)) / sum(m1.test)
```

```
## [1] 0.7058824
```

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

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



```
# - Almost 2/3 of Adelie penguins in the testing data are misclassified
# - All of the Chinstrap and Gentoo penguins in the testing data are correctly classified!
```

```
# Make predictions for test observations based on tree2
test_preds_2 <- predict(tree2, newdata=test, type="class")
head(test_preds_2)
```

```
##      1      2      3      4      5      6
## Adelie Adelie Adelie Adelie Adelie Adelie
## Levels: Adelie Chinstrap Gentoo
```

```
m2.test <- table(test_preds_2, test$species)
m2.test
```

```
##
## test_preds_2 Adelie Chinstrap Gentoo
##      Adelie      29      1      0
##      Chinstrap    2     11      0
##      Gentoo       0      0     25
```

```
# What is the accuracy for tree1 based on testing data?
sum(diag(m2.test)) / sum(m2.test)
```

```
## [1] 0.9558824
```

```
# Make predictions for test observations based on tree3
test_preds_3 <- predict(tree3, newdata=test, type="class")
head(test_preds_3)
```

```
##      1      2      3      4      5      6
## Adelie Adelie Adelie Adelie Adelie Adelie
## Levels: Adelie Chinstrap Gentoo
```

```
m3.test <- table(test_preds_3, test$species)
m3.test
```

```
##
## test_preds_3 Adelie Chinstrap Gentoo
##      Adelie      29      1      0
##      Chinstrap    2     11      0
##      Gentoo       0      0     25
```

```
# What is the accuracy for tree1 based on testing data?
sum(diag(m3.test)) / sum(m3.test)
```

```
## [1] 0.9558824
```

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

```
# They are the same!
```

```
# If we look at the trees more closely, we see that although one of the splits involves a different variable,
```

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

```
# In this case, both trees have very similar complexity (similar number of splits/terminal nodes), but tree1 is simpler.
```