

FISH 6003: Week 1 Example

Brett Favaro

January 14, 2018

R Markdown

I recommend using R Markdown when completing assignments. Information is available online at <http://rmarkdown.rstudio.com>.

In R Markdown you can embed R code. Here is a version of our Week 1 data exploration, but formatted nicely in Markdown.

Loading and Verifying data

Load data and packages. Notice that in Markdown I have to use two dots, not one.

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Conflicts with tidy packages -----

## filter(): dplyr, stats
## lag():    dplyr, stats

IntroData <- read.csv("../data/6003-1-introdata.csv")
```

Check whether it loaded correctly.

```
head(IntroData)
```

```
##   fishid      species length_cm
## 1      1 atlantic_salmon      41
## 2      2 atlantic_salmon      38
## 3      3 atlantic_salmon      43
## 4      4 atlantic_salmon      46
## 5      5 atlantic_salmon      31
## 6      6 atlantic_salmon      50
```

```
str(IntroData)
```

```
## 'data.frame':   20 obs. of  3 variables:
## $ fishid   : int  1 2 3 4 5 6 7 8 9 10 ...
## $ species  : Factor w/ 2 levels "atlantic_salmon",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ length_cm: int  41 38 43 46 31 50 49 47 45 46 ...
```

The variables are:

- fishid: One identifier per fish
- species: Common name of the species being measured
- length_cm: total length of fish being measured.

Let's verify that each one looks like it's error-free.

fishid

This should be a number with one unique value per fish. `str()` above told us it's an integer.

Let's look at the range of values, and check for duplicates:

```
range(IntroData$fishid)
```

```
## [1] 1 20
```

```
IntroData %>% # Do stuff to IntroData
  group_by(fishid) %>% # 1. Group by the fish id
  filter(n() > 1) # 2. Return anything that occurs more than once
```

```
## # A tibble: 0 x 3
```

```
## # Groups:   fishid [0]
```

```
## # ... with 3 variables: fishid <int>, species <fctr>, length_cm <int>
```

No duplicates. All good.

species

There should just be two levels of species. Is that true?

```
levels(IntroData$species)
```

```
## [1] "atlantic_salmon" "porbeagle_shark"
```

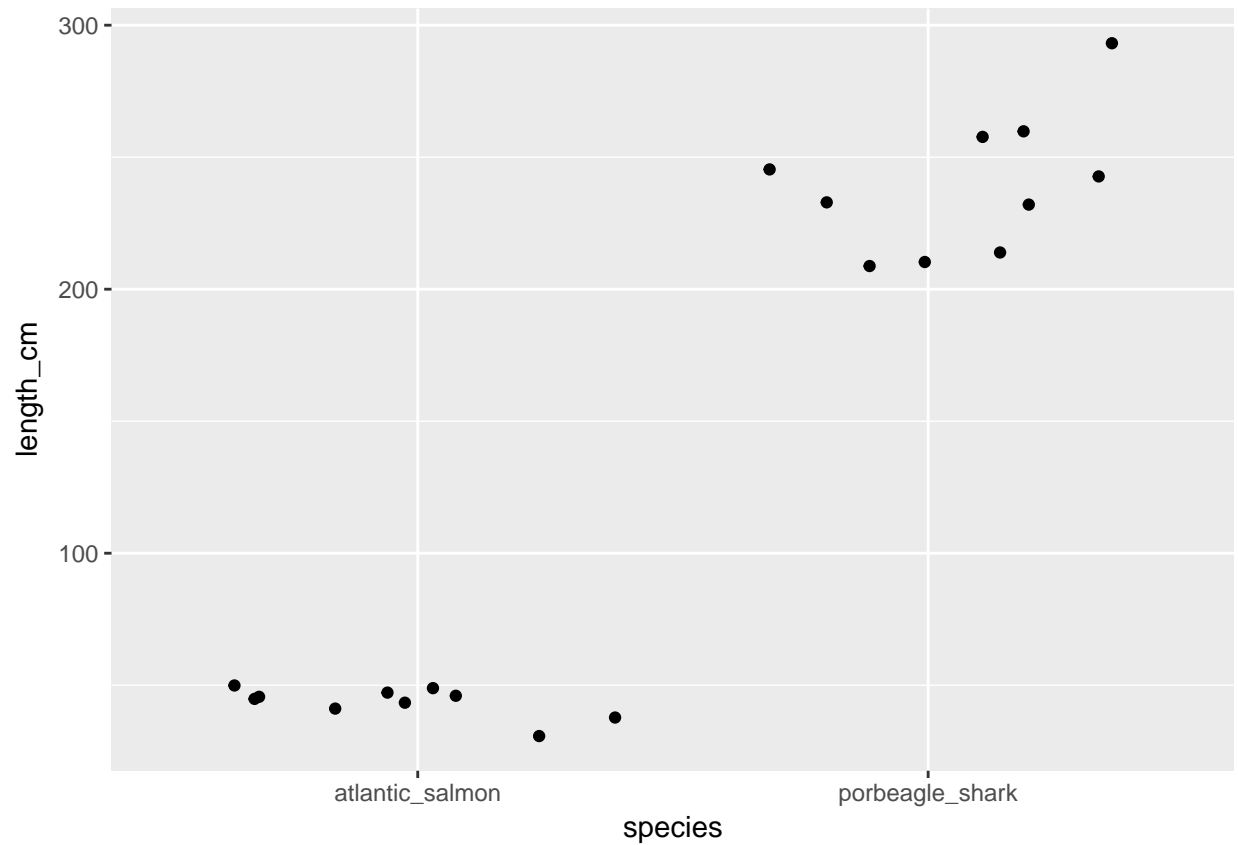
Yep, all good. Moving on.

length_cm

Let's make sure the total lengths are okay. Plot:

```
p <- ggplot(IntroData, aes(x = species, y = length_cm))
```

```
p + geom_jitter()
```



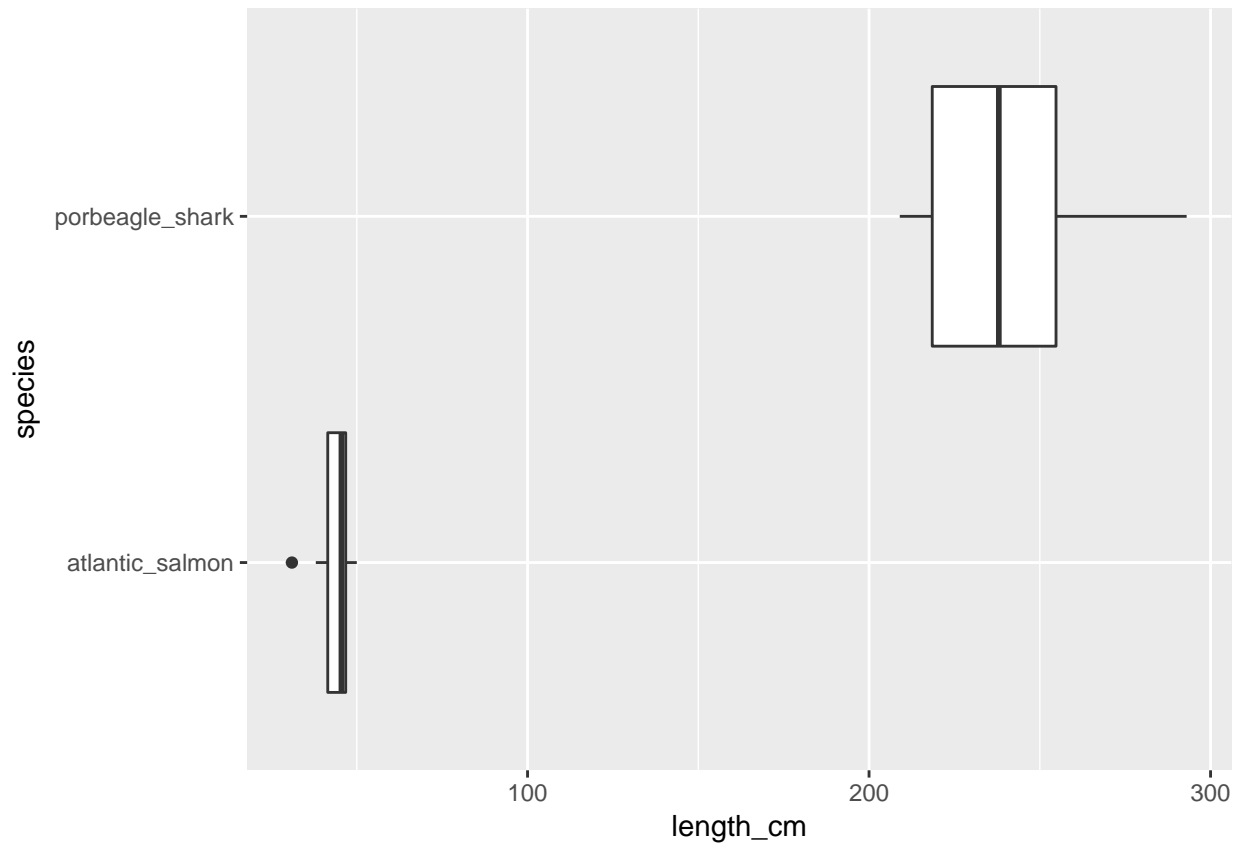
Are these values reasonable? If you collected the data yourself you'd know - but you didn't do that here (I gave you the data myself).

One option is to trust me. But if you're suspicious, go to <http://fishbase.org/> and look up these fish species. Never hurts to check!

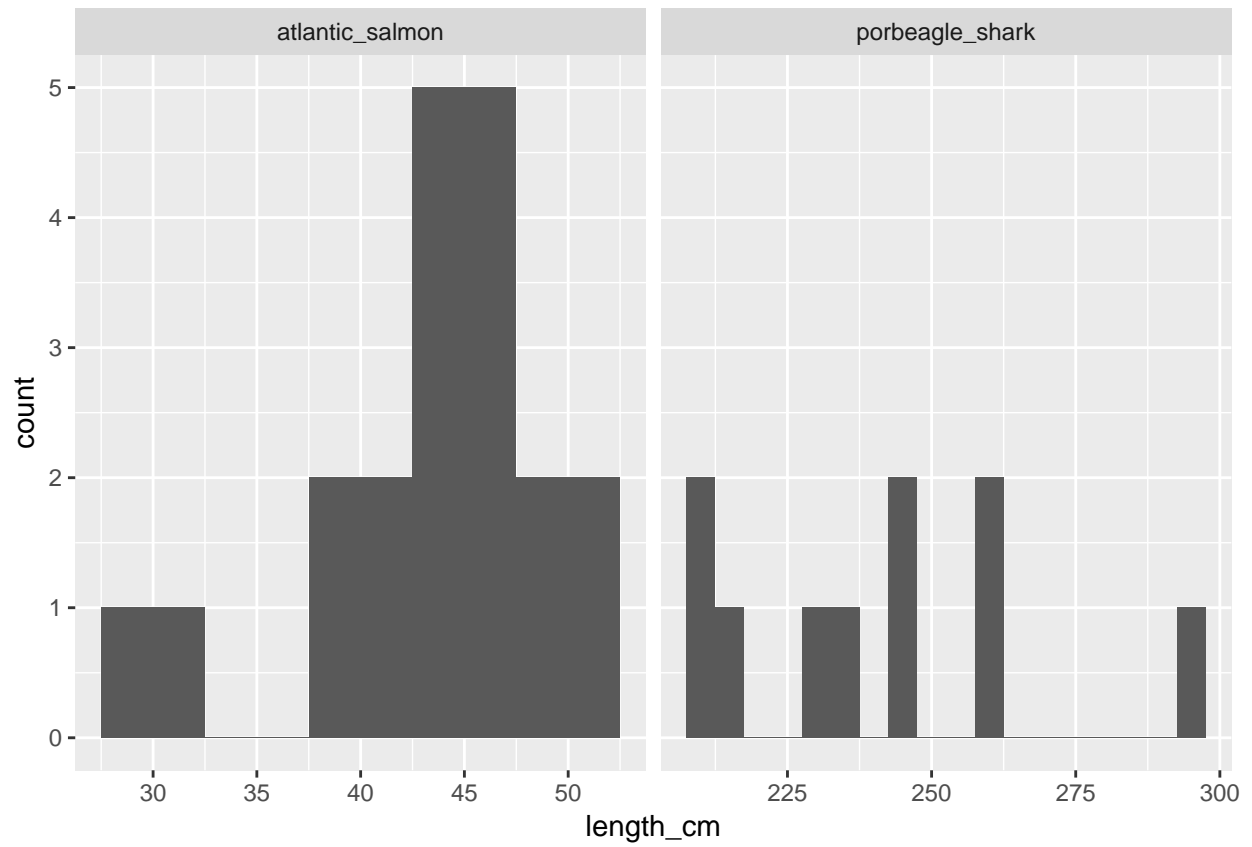
Analysis

Not much we can do from these data, aside from some very basic visualizations.

Let's make a boxplot



Nice. Now, a histogram:



Here, I've used `echo=FALSE` so you don't see my R code. You can check the Rmd file for it if you want.