Abalone Exploratory Data Analysis

Ryan Heslin

April 19, 2022

```
abalone <- read.csv(here::here("data", "abalone_raw.csv"))
library(ggplot2)
```

The data for this project relate to measurements of abalone, a species of aquatic snail sometimes eaten as a delicacy. It may be downloaded here. There are 4177 observations of 9 rows. The response, variable, **sex**, has three levels: "M", "F", and "I", corresponding to male, female, and infant. Available predictors correspond to measurements of size and weight, in addition to the number of rings. As with trees, counting rings approximates age.

```
abalone_long <- abalone |> tidyr::pivot_longer(ends_with("weight"), names_to = "Measure", values_to = "
```

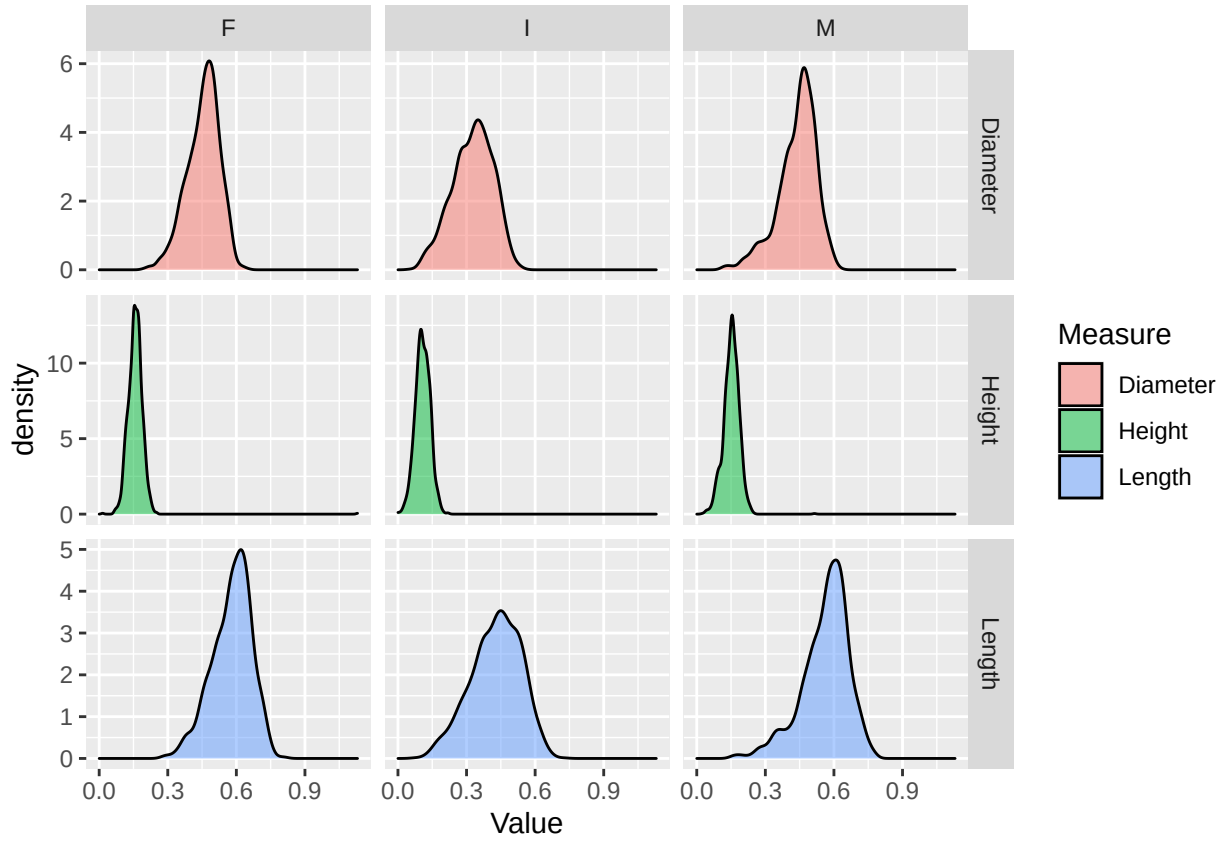
The three classes of `Sex` are about evenly distributed.

```
table(abalone$Sex)
```

```
   F    I    M
1307 1342 1528
```

Distributions of dimension measures are quite similar across the classes of `sex`.

```
abalone |>
  tidyr::pivot_longer(c(Length, Diameter, Height), names_to = "Measure", values_to = "Value") |>
  ggplot(aes(x = Value, fill = Measure)) +
  geom_density(alpha = .5) +
  facet_grid(rows = vars(Measure), vars(Sex), scales = "free_y")
```



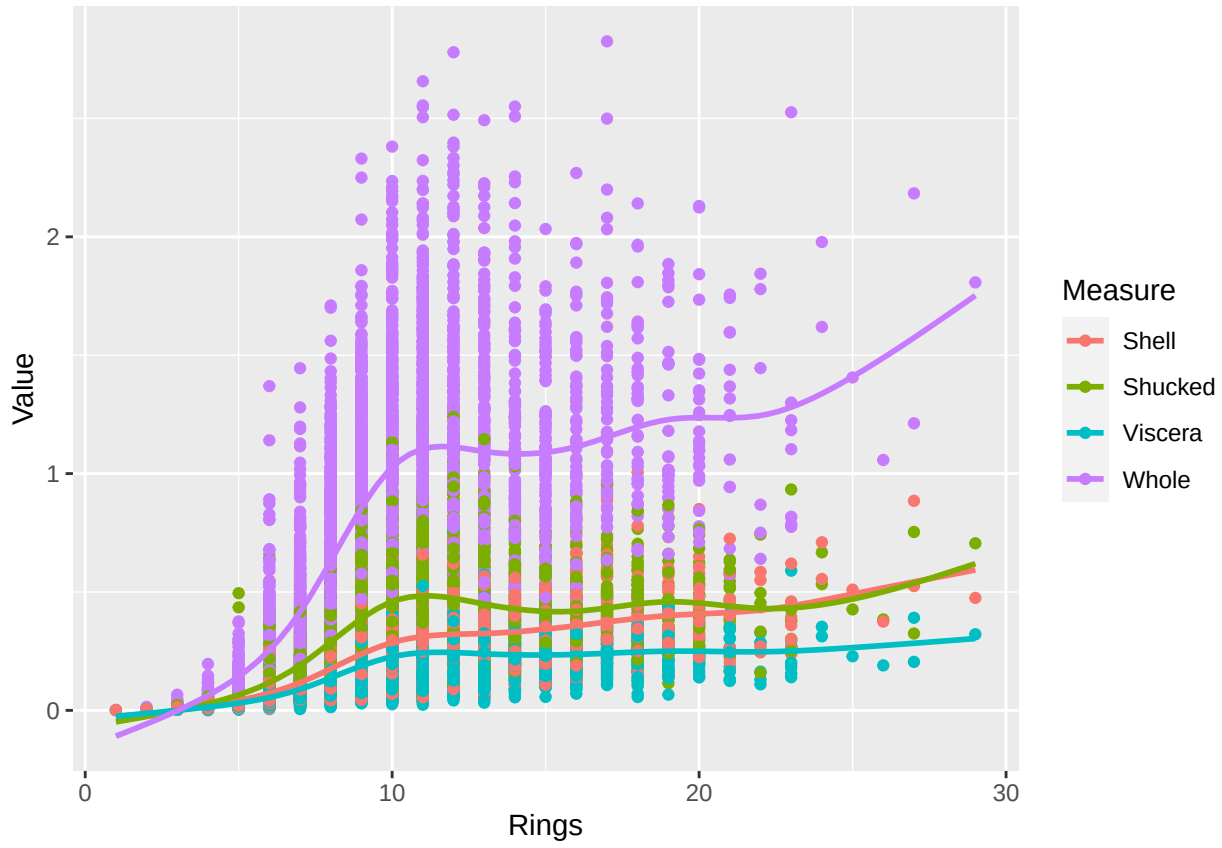
The three separate weight variables don't quite sum to total weight, suggesting measurement error.

```
summary(with(abalone, Whole.weight - Shucked.weight - Viscera.weight - Shell.weight))
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.
-0.44750  0.01800  0.03700  0.04995  0.06800
    Max.
 0.60800
```

It seems that weight increases nonlinearly with number of rings. This makes sense, as many organisms grow at a nonlinear rate.

```
ggplot(abalone_long, aes(x = Rings, y = Value, color = Measure)) +
  geom_point() +
  geom_smooth(se = FALSE)
```



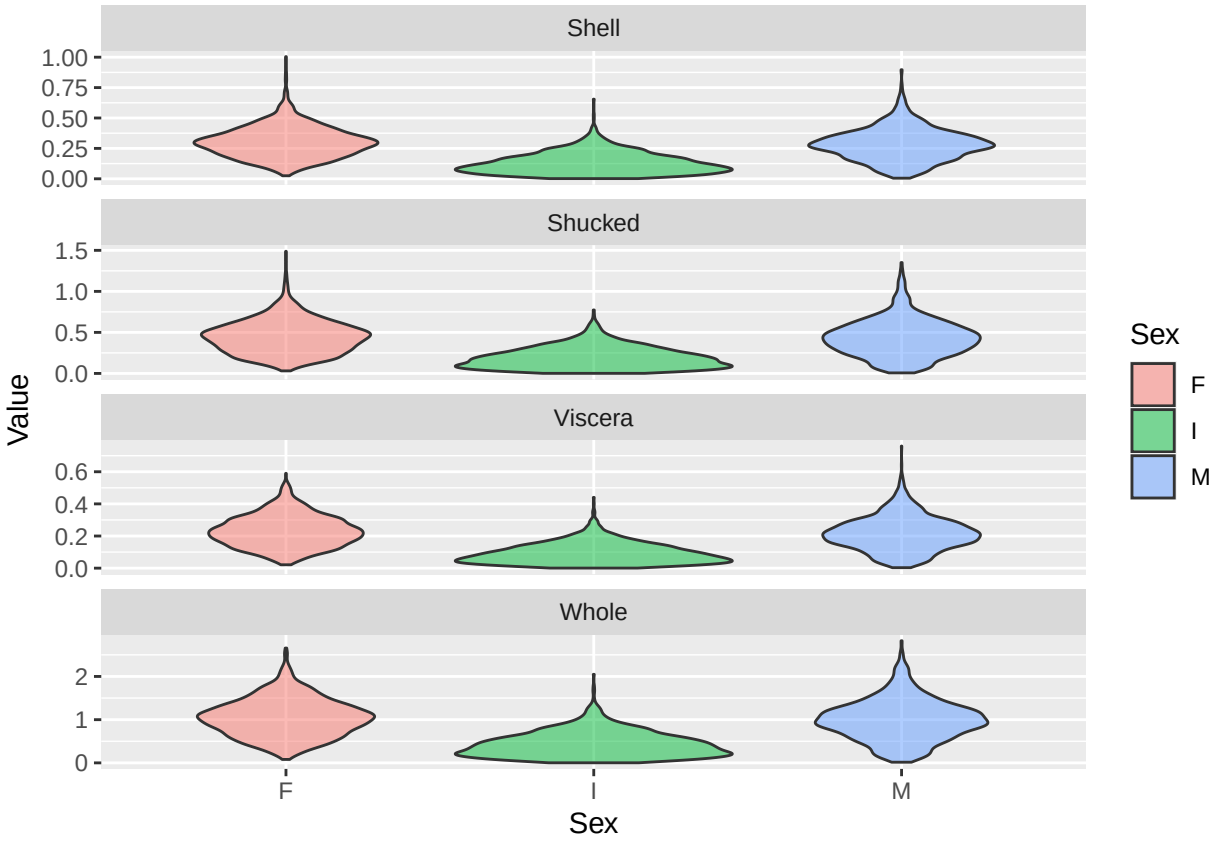
Infants have substantially fewer rings than adults, but adults differ only a little

```
with(abalone, tapply(Rings, Sex, mean))
```

```
       F        I        M
11.129304  7.890462 10.705497
```

Adults of both sexes have very similar weight distributions, but infants weigh notably less, as would be expected

```
ggplot(abalone_long, aes(x = Sex, fill = Sex, y = Value)) +
  geom_violin(alpha = .5) +
  facet_wrap(~Measure, ncol = 1, scales = "free_y")
```



Applying K-means clustering with $K = 3$ does not result in well-separated assignments. This is consistent with the major variation occurring between infants and adults, of both sexes, not males and females.

```
set.seed(1)
clusters <- abalone[, -match("Sex", colnames(abalone))] |>
  scale() |>
  kmeans(centers = 3) |>
  getElement("cluster")

table(clusters, abalone[["Sex"]])
```

```
clusters   F   I   M
       1 586  35 606
       2 596 467 710
       3 125 840 212
```