NRSG 741 - Homework 2 - ANSWER KEY

Melinda Higgins

03/25/2019

INSTRUCTIONS

- Use this Rmarkdown file `N741Homework02.Rmd` to get started.
- Change the author to YOUR NAME
- Note: This Rmarkdown file has one R code chunk at the top that reads in the dataset and loads the R packages you will need.
- After each question below, insert an R code chunk to enter the R code needed to answer that question. Do this for each question.
- Outside of the R code chunk, type in any text needed to provide explanation or answer the questions further.

Due Date is 13 February 2019

This homework is meant to further your `dplyr` and `ggplot2` skills.

Abalones Dataset from UCI Repository

For this homework, you will keep working with the `abalone` dataset from the UCI data repository at https://archive.ics.uci.edu/ml/datasets/abalone.

Use tools within the `dplyr` package as much as possible to answer the following questions.

Question 1: What kind of R object is the `abalone` dataset?

ANSWER KEY

You could use either the `class()` or `str()` functions to give you details about the `abalone` dataset object.

```
# insert R code here to answer question 1

class(abalone)
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
str(abalone)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 4177 obs. of  9 variables:
##  $ sex          : chr  "M" "M" "F" "M" ...
##  $ length       : num  0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 ...
##  $ diameter     : num  0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 ...
##  $ height       : num  0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 ...
##  $ wholeWeight  : num  0.514 0.226 0.677 0.516 0.205 ...
##  $ shuckedWeight: num  0.2245 0.0995 0.2565 0.2155 0.0895 ...
##  $ visceraWeight: num  0.101 0.0485 0.1415 0.114 0.0395 ...
##  $ shellWeight  : num  0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 ...
```

```
##  $ rings        : num  15 7 9 10 7 8 20 16 9 19 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   X1 = col_character(),
##   ..   X2 = col_double(),
##   ..   X3 = col_double(),
##   ..   X4 = col_double(),
##   ..   X5 = col_double(),
##   ..   X6 = col_double(),
##   ..   X7 = col_double(),
##   ..   X8 = col_double(),
##   ..   X9 = col_double()
##   .. )
```

ANSWER KEY

The `abalone` dataset is read in as a `data.frame`, using the `read_csv()` function from the `readr` package which is part of the `tidyverse`, actually makes it a `"spec_tbl_df" "tbl_df" "tbl" "data.frame"` - a tibble data frame. You can learn more at https://www.tidyverse.org/articles/2018/12/readr-1-3-1/.

Question 2: How many observations are in the `abalone` dataset?

ANSWER KEY

To answer this question, you can use either the `str()` or `dim()` functions.

```
dim(abalone)
```

```
## [1] 4177    9
```

ANSWER KEY

Based on either the `str()` output shown above or the `dim()` results, there are 4177 observations in the `abalone` dataset.

Question 3: For shucked weight, how many abalones weigh more than 0.8 grams?

ANSWER KEY

Using the `filter()` function from the `dplyr` package is useful for extracting cases (observations or rows) that meet the specified criteria defined inside the `filter()` function. Only rows for which the filter is TRUE are retained.

```
abalone %>%
  filter(shuckedWeight > 0.8) %>%
  dim()
```

```
## [1] 148   9
```

There are 148 abalones with a shucked weight > 0.8 grams.

Question 4: How many abalones have shucked weights larger than their whole weight?

(HINT: create a new variable using mutate and then filter)

ANSWER KEY

This problem could have been solved using either `mutate()` or the `filter()` function. Both approached should yield the same answer.

```
abalone %>%
  mutate(shuckedHigh = shuckedWeight > wholeWeight) %>%
  filter(shuckedHigh == TRUE) %>%
  dim()
```

```
## [1]  4 10
```

```
# alternate approach without mutate

abalone %>%
  filter(shuckedWeight > wholeWeight) %>%
  dim()
```

```
## [1] 4 9
```

There are 4 abalones with shucked weight greater than their whole weight which should not be correct. These abalones have measurement errors.

Create a subset containing only infants `sex == "I"`

Question 5: How many infants are in this subset?

ANSWER KEY

Create the subset first and then find the dimensions to get number of rows.

```
# Create subset
abaloneI <- abalone %>%
  filter(sex == "I")

# Find dimensions
dim(abaloneI)
```

```
## [1] 1342    9
```

There are 1342 infant abalones in this dataset.

Show off your `dplyr` skills with `group_by()`

Question 6: What is the average whole weight for each abalone sex (get whole weight means for females "F", males "M" and infants "I" separately)?

ANSWER KEY

You can use either `summarise()` or `summarise_all()` functions from `dplyr` package.

```
abalone %>%
  group_by(sex) %>%
  summarise(meanwt = mean(wholeWeight))
```

```
## # A tibble: 3 x 2
##   sex   meanwt
##   <chr>  <dbl>
## 1 F      1.05
## 2 I      0.431
## 3 M      0.991
```

```
abalone %>%
  group_by(sex) %>%
  select(wholeWeight) %>%
  summarise_all(mean)
```

```
## Adding missing grouping variables: `sex`
```

```
## # A tibble: 3 x 2
##   sex   wholeWeight
##   <chr>       <dbl>
## 1 F           1.05
## 2 I           0.431
## 3 M           0.991
```

Question 7: Get the means for the abalone length and height by sex?

ANSWER KEY

You can use either `summarise()` or `summarise_all()` functions from `dplyr` package. This is very similar to problem above, notice only the variable names get updated.

```
abalone %>%
  group_by(sex) %>%
  summarise(meanlt = mean(length),
            meanht = mean(height))
```

```
## # A tibble: 3 x 3
##   sex   meanlt meanht
##   <chr>  <dbl>  <dbl>
## 1 F      0.579  0.158
## 2 I      0.428  0.108
## 3 M      0.561  0.151
```

```
abalone %>%
  group_by(sex) %>%
  select(length, height) %>%
  summarise_all(mean)
```

```
## Adding missing grouping variables: `sex`
```

```
## # A tibble: 3 x 3
##   sex   length height
##   <chr>  <dbl>  <dbl>
## 1 F      0.579  0.158
## 2 I      0.428  0.108
## 3 M      0.561  0.151
```

Test your graphing skills using ggplot2

Using the `abalone` dataset, create the following graphics/figures using `ggplot()` and associated `geom_xxx()` functions.

Question 8: Create a histogram of abalone whole weight

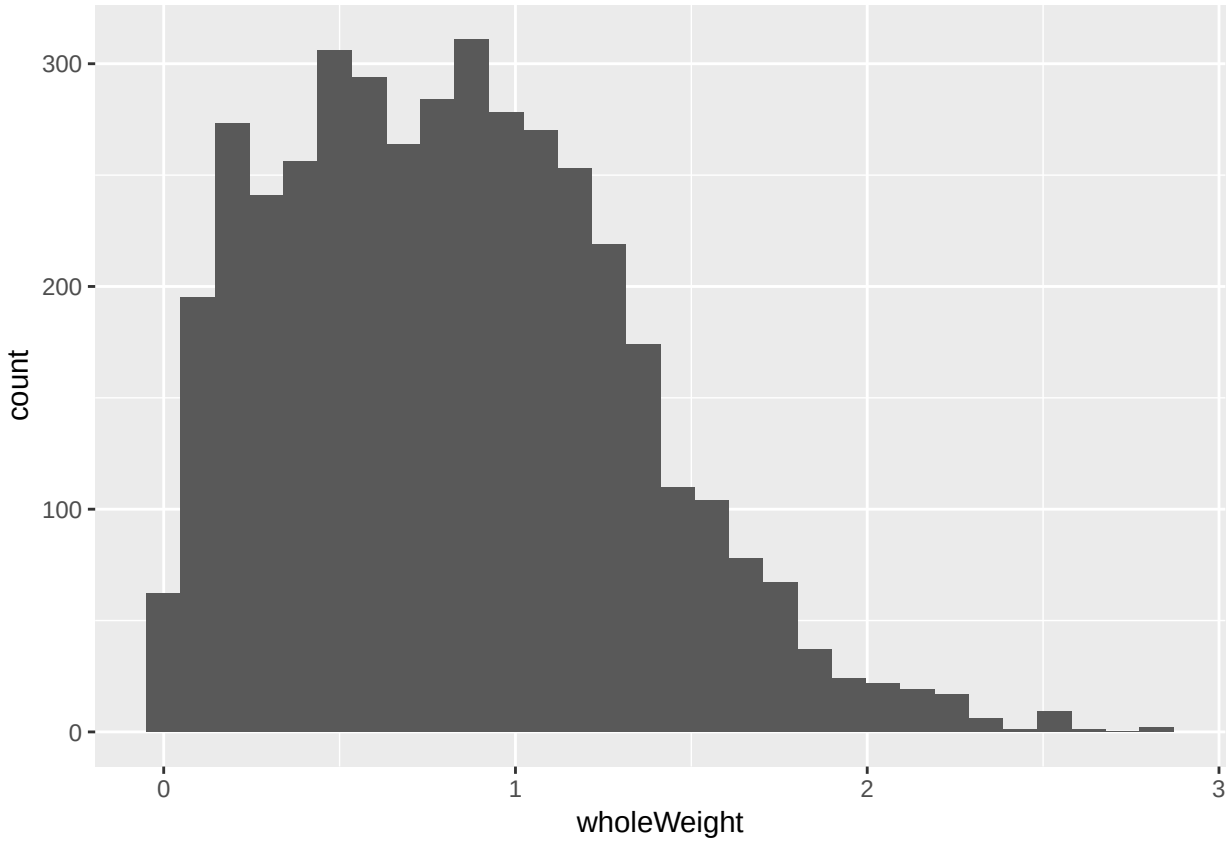
What do you notice about the distribution (any outliers or skewness)?

ANSWER KEY

You want to use the `geom_histogram()` function from the `ggplot2` package. In the initial `ggplot()` step, you only have to define one aesthetic (`aes`) for `wholeWeight`.

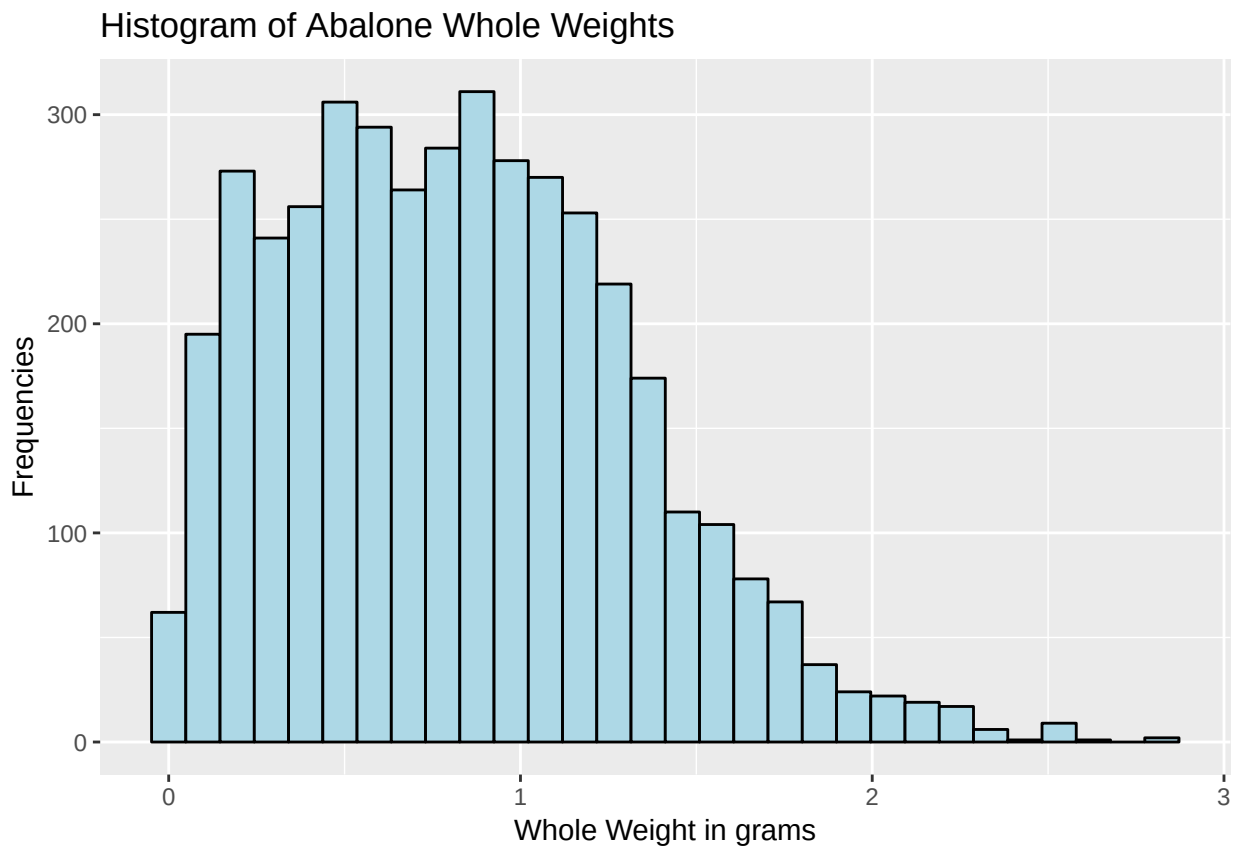
```
# simple histogram
ggplot(abalone, aes(x=wholeWeight)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# optional - add colors and labels and a title
ggplot(abalone, aes(x=wholeWeight)) +
  geom_histogram(color = "black",
                 fill = "light blue") +
  xlab("Whole Weight in grams") +
  ylab("Frequencies") +
  ggtitle("Histogram of Abalone Whole Weights")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There are a few large abalones with weights above 2.5 grams.

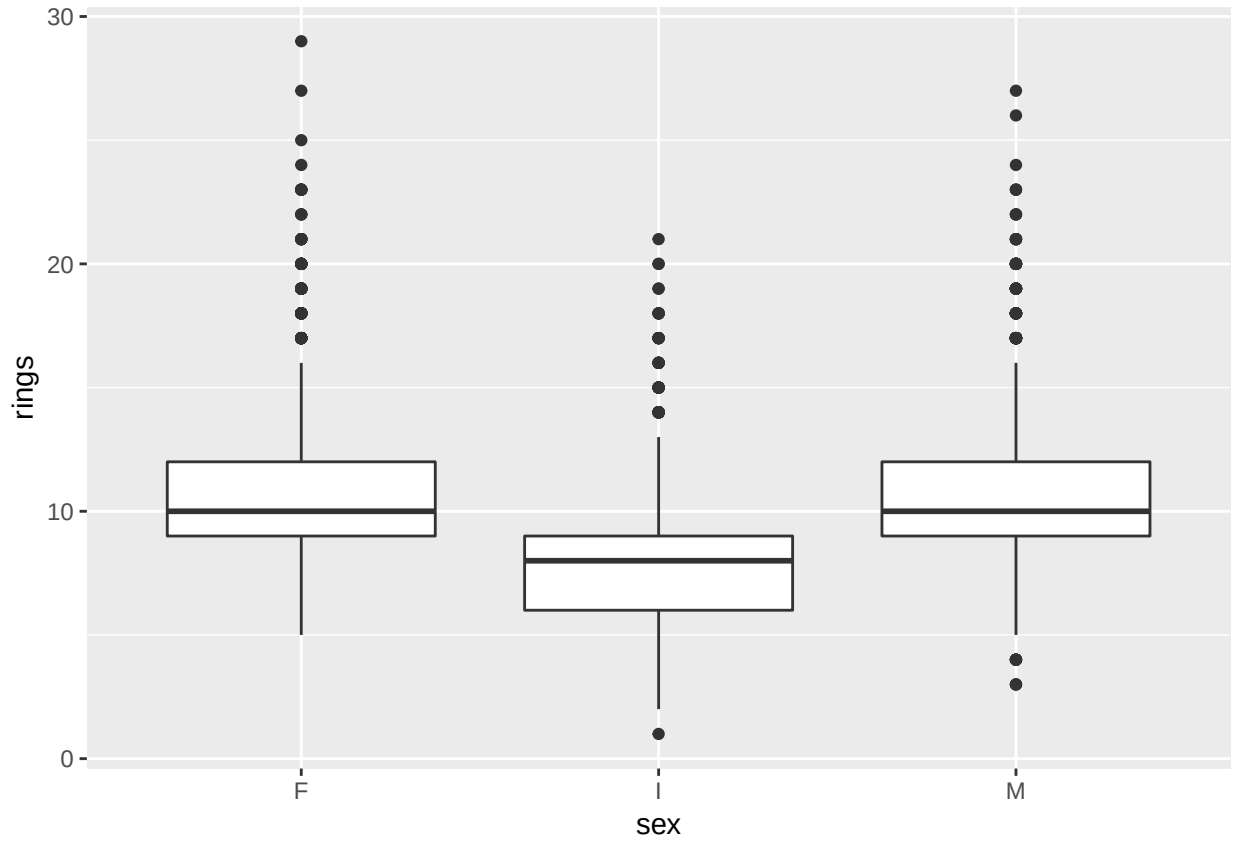
Question 9: Create side-by-side boxplots of the number of rings by gender

HINT use `geom_boxplot` *with* $x = sex$ *and* $y = rings$

ANSWER KEY

Use similar approach to above, but now you have two aesthetics (`aes`) instead of just one for the histogram above. The two aesthetics (`aes`) are `sex` and `rings`. You need `geom_boxplot()` to draw the boxplots.

```
ggplot(abalone, aes(x=sex, y=rings)) +
  geom_boxplot()
```



Question 10: Create a scatterplot of the whole weight on the X axis and shucked weight on the Y axis and color the points by sex

ANSWER KEY

Scatterplots also need two aesthetics (`aes`) - in this case are `wholeWeight` for "x" and `shuckedWeight` for "y". You also need `geom_point()` and color the points by `sex` using `aes(color)` inside `geom_point()`.

```
ggplot(abalone, aes(x=wholeWeight, y=shuckedWeight)) +
  geom_point(aes(color = sex))
```


1.5
1.0
0.5
0.0
shuckedWeight
0
1
2
wholeWeight
sex
F
I
M