NRSG 741 - Homework 2 - Exploratory Data Analysis

Alexander Maillis

2/11/2020

INSTRUCTIONS

- Use this Rmarkdown file `N741Spring2020_Homework02.Rmd` to get started.
- Change the author to YOUR NAME
- Change the date
- 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.

Note: BEFORE you Knit your document, be sure to comment out any code that is not fully running yet by adding a `#` at the beginning of that line of code.

Note: All you need to do is correctly fill in the blanks `___` in the code chunks below.

Goal of Homework 2

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

Modify R code chunks

In each of the R code chunks below, scaffolding is provided. Everywhere you see 3 underscores _____ , you will need to fill in the appropriate code, variable name, function name, etc.

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?

```
# insert R code here to answer question 1
# HINT: The name of the dataset is abalone

class(abalone)
```

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

Data Frame

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

```
# HINT: there are multiple ways to answer this - pick one
dim(___)
str(___)
glimpse(___)
nrow(abalone)
```

```
## Error: <text>:2:5: unexpected input
## 1: # HINT: there are multiple ways to answer this - pick one
## 2: dim(_
##         ^
```

4177

Question 3: For diameter, how many abalones have diameters less than 0.5mm?

```r
# the variable name is diameter
abalone %>%
 filter(diameter< 0.5) %>%
 nrow()
```

```
## [1] 3388
```

3388

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

NOTE: There should be NO measurements where the shucked weight is > whole weight. If there are some these are probably data entry errors in this dataset.

```r
# HINT: Use a logical expression inside a filter step
# HINT: Check the spelling and case for the
# variable names for shucked weight and whole weight
abalone %>%
 filter(shuckedWeight> wholeWeight) %>%
 nrow()
```

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

4

Create a subset containing only infants `sex == "I"`. Call this new dataset `infants`

```r
# HINT: Put the logical statement inside the filter() function
# Dont forget to use the assign operator <- to create the infants object
infants <- abalone %>%
  filter(sex == "I")
```

Question 5: How many infants are in this subset?

```r
# Hint: see code in question 2 above
# pick the function you prefer to answer this question
nrow(infants)
```

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

1342

Show off your `dplyr` skills with `group_by()` - we didn't get a chance to fully explore `group_by()` in class but it is added in the examples below to help you answer these questions.

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)?

```r
# Hint: put the variables used in the select statement
# and in the summarise() statement. Remember to put
# in a name for the output of the mean() function
# something like meanwt
abalone %>%
  select(sex, wholeWeight) %>%
  group_by(sex) %>%
  summarise(mean1 = mean(wholeWeight, na.rm=TRUE))
```

```
## # A tibble: 3 x 2
##   sex   mean1
##   <chr> <dbl>
## 1 F     1.05
## 2 I     0.431
## 3 M     0.991
```

sex mean1

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?

```r
# Hint: put variable names in the select statement
# put the function name for the mean in the
# summarise_all() function
abalone %>%
  select(sex, length, height) %>%
  group_by(sex) %>%
  summarise_all(mean, na.rm=TRUE)
```

```
## # 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
```

sex length height

1 F 0.579 0.158

2 I 0.428 0.108

3 M 0.561 0.151

Test your graphing skills using ggplot2

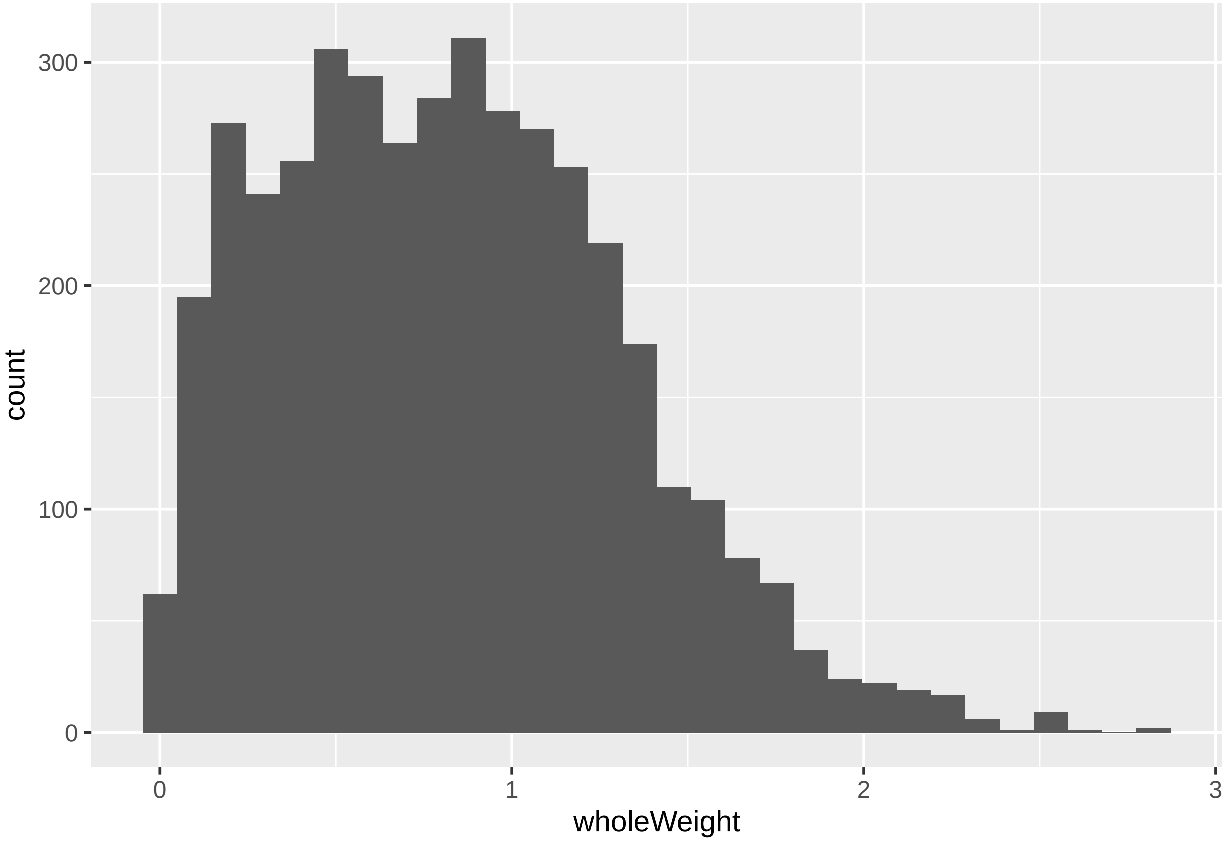
Using the `abalone` dataset, create the following graphics/figures using `ggplot()` and associated `geom_xxx()` functions.

Question 8: Create a histogram of abalone whole weight

BONUS: Outline the histogram bars with a black line and fill the histogram bars with a green color

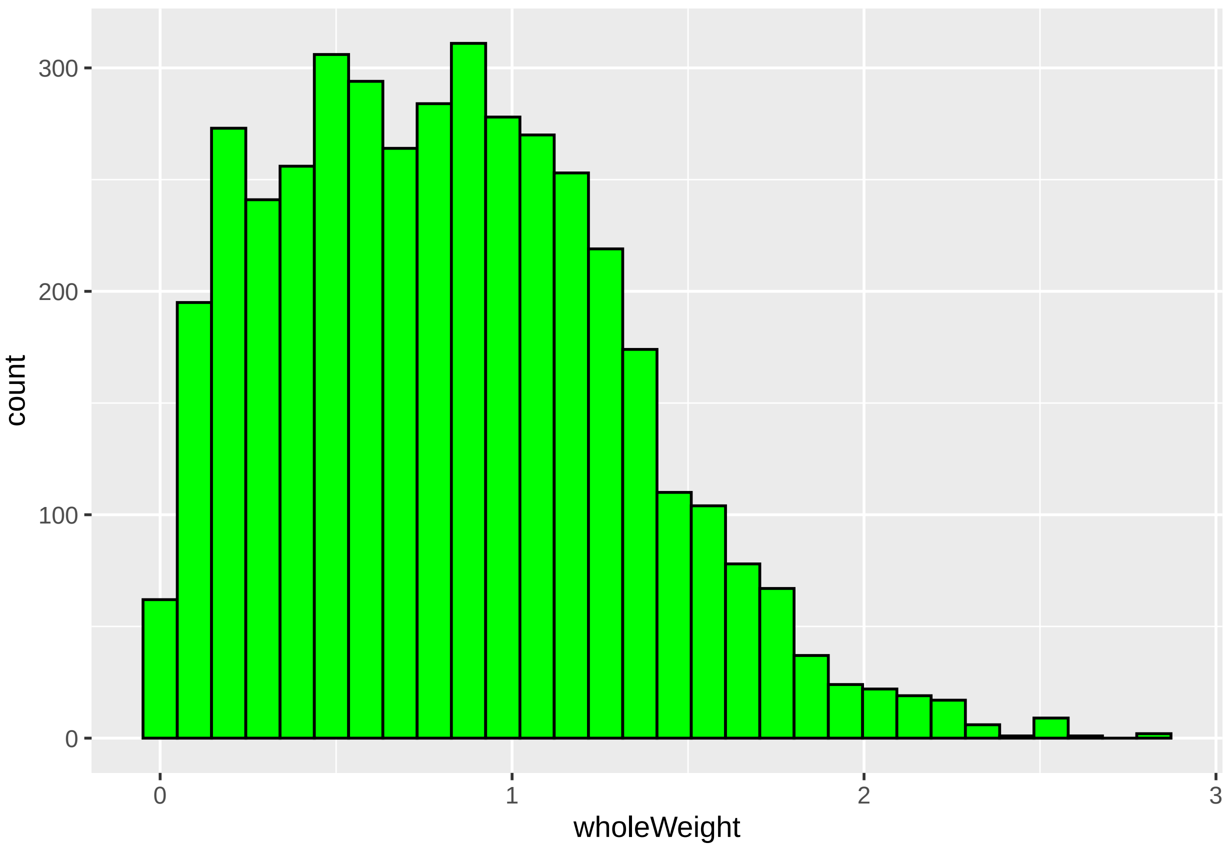
```
# Hint: the first option in the ggplot() function
# is the name of the dataset
# the variable name is put inside aes()
ggplot(abalone, aes(x=wholeWeight)) +
  geom_histogram()
```

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



```
# BONUS:
ggplot(abalone, aes(x=wholeWeight)) +
  geom_histogram(color = "black", fill = "green")
```

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



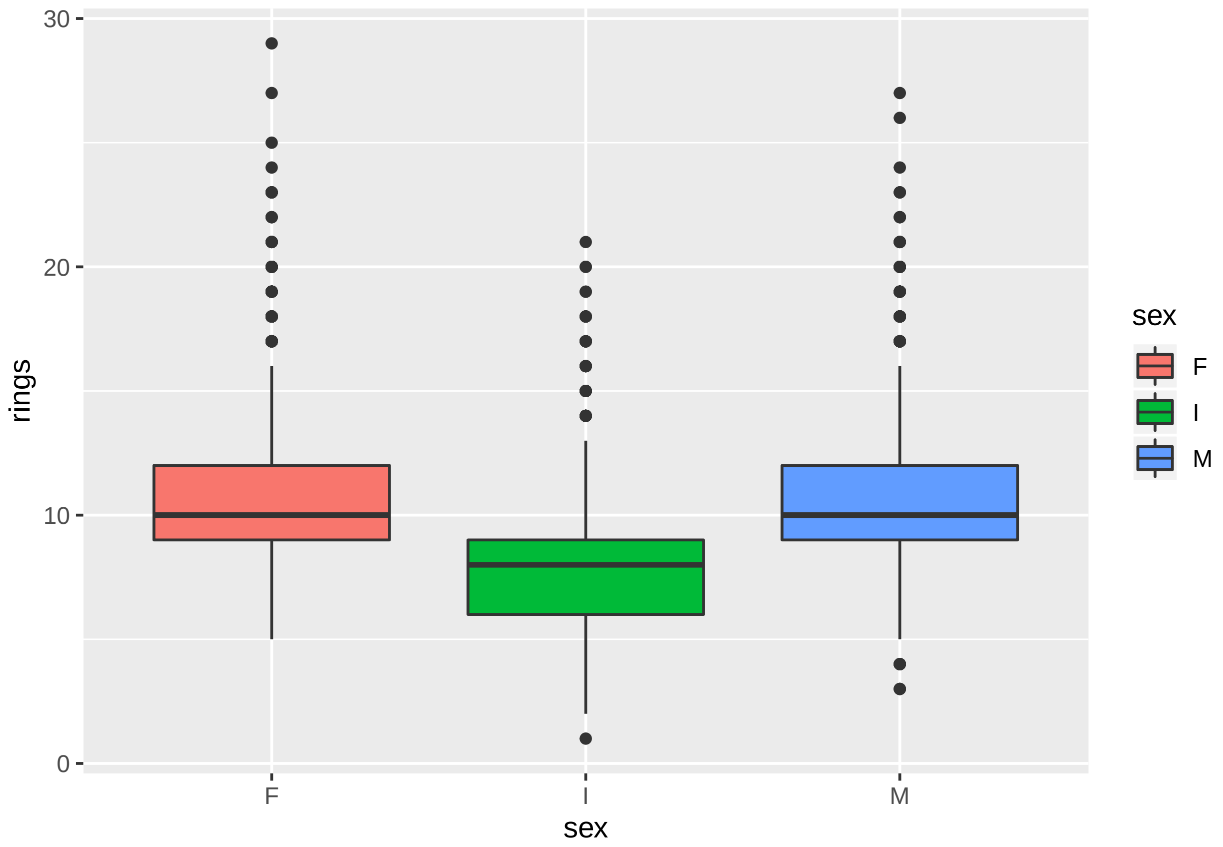
What do you notice about the distribution (any outliers or skewness)?

Right skewed distribution from the outliers.

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

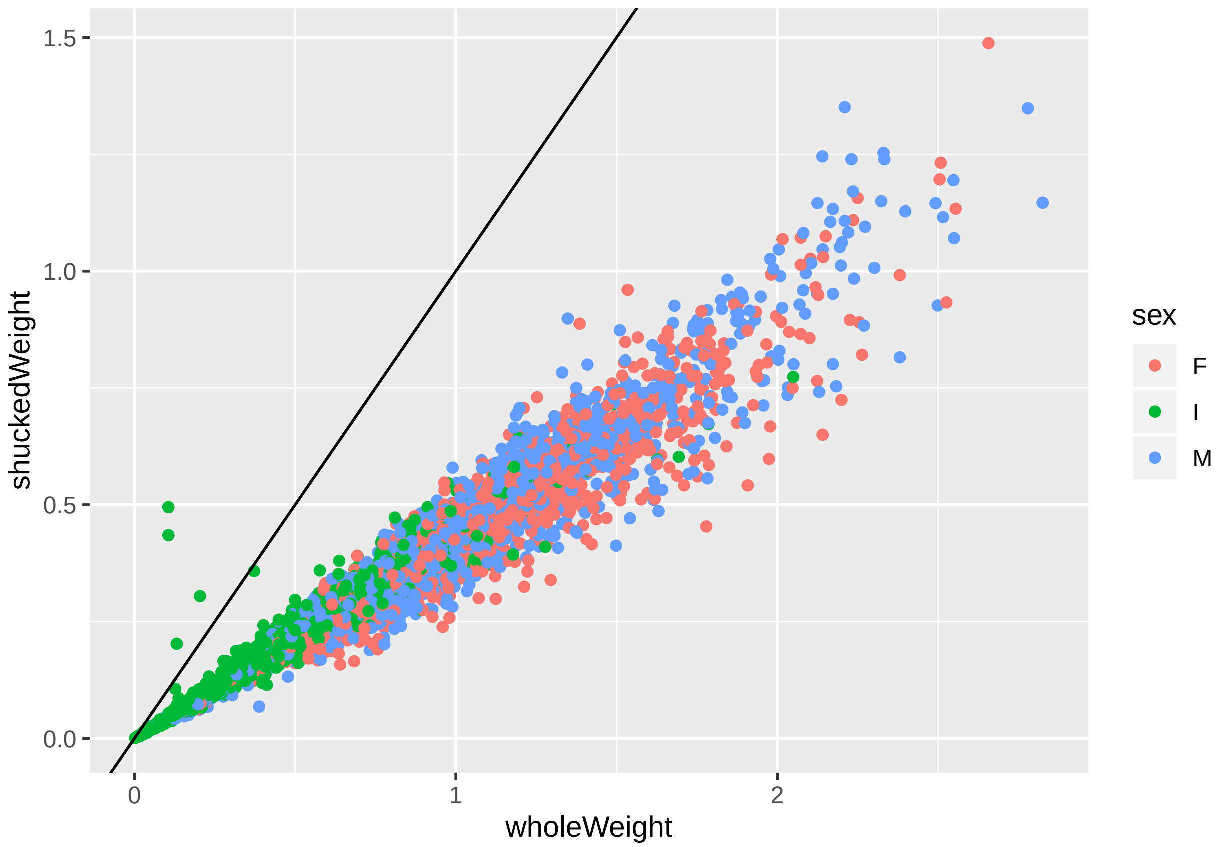
HINT use `geom_boxplot` with x = sex and y = rings

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

```r
ggplot(abalone, aes(x=wholeWeight, y=shuckedWeight, color=sex)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1)
```



Can you see which abalones have shucked weights > whole weights which should not happen? Look at the y=x reference line. What sex are the abalones with the incorrect weights?

Final Instructions

- KNIT this RMD file to PDF (or to HTML or DOC and save as PDF)
- Upload your PDF document to Canvas