Practice Solutions to Intro to R and Rstudio for EDA - Part 2

Jessica Minnier, PhD & Meike Niederhausen, PhD

OCTRI Biostatistics, Epidemiology, Research & Design (BERD) Workshop

2020/09/17

Practice 3

1. Continue adding code chunks to your Rmd (or, start a new one! But remember to load the libraries and data at the top.)

2. How many different years are in the data? (Hint: use `tabyl()` or `n_distinct()`)

3. Count the number of penguins measured each year.

4. Calculate the median body mass by each species and sex subgroup. Use `summarize()` and `group_by()` to do this.

5. Create a 2x2 table of number of penguins measured in each year by each island.

Practice 3 Answers

2. How many different years are in the data? (Hint: use `tabyl()` or `n_distinct()`)

Option 1:

```
penguins %>%
  summarize(n_distinct(year))
```

```
# A tibble: 1 x 1
  `n_distinct(year)`
               <int>
1                  3
```

Option 2:

```
penguins %>% tabyl(year)
```

```
 year   n   percent
 2007 109 0.3187135
 2008 114 0.3333333
 2009 119 0.3479532
```

```
penguins %>% tabyl(year) %>% nrow
```

```
[1] 3
```

3. Count the number of penguins measured each year.

Option 1:

```
penguins %>% count(year)
```

```
# A tibble: 3 x 2
   year     n
  <dbl> <int>
1  2007   109
2  2008   114
3  2009   119
```

Option 2:

```
penguins %>% tabyl(year)
```

```
 year   n   percent
 2007 109 0.3187135
 2008 114 0.3333333
 2009 119 0.3479532
```

4. Calculate the median body mass by each species and sex subgroup. Use `summarize()` and `group_by()` to do this.

```
penguins %>%
  group_by(species, sex) %>%
  summarize(median(body_mass_g))
```

```
# A tibble: 8 x 3
# Groups:   species [3]
  species   sex    `median(body_mass_g)`
  <chr>     <chr>                  <dbl>
1 Adelie    female                 3400
2 Adelie    male                   4000
3 Adelie    <NA>                   3475
4 Chinstrap female                 3550
5 Chinstrap male                   3950
6 Gentoo    female                 4700
7 Gentoo    male                   5500
8 Gentoo    <NA>                   4688.
```

5. Create a 2x2 table of number of penguins measured in each year by each island.

```
penguins %>% tabyl(island, year)
```

```
    island 2007 2008 2009
    Biscoe   44   64   59
     Dream   46   34   44
 Torgersen   19   16   16
```

Practice 4 (old practice 2)

1. Create a new script and save it as `Practice2.R`
2. Create data frames for males and females separately.
3. Do males and females have similar BMIs? Weights? Compares means, standard deviations, range, and boxplots.
4. Plot BMI vs. weight for each gender separately. Do they have similar relationships?
5. Are males or females more likely to be bullied in the past 12 months? Calculate the percentage bullied for each gender.

Practice 4 Answers

#2 Create data frames for males and females separately.

```
boys <- mydata[mydata$sex == "Male", ]
dim(boys)
```

```
[1]  8 10
```

```
girls <- mydata[mydata$sex == "Female", ]
dim(girls)
```

```
[1] 12 10
```

Check number of boys & girls:

```
summary(mydata$sex)
```

```
   Length     Class      Mode
       20 character character
```

#3 Do males and females have similar BMIs? Weights? Compares means, standard deviations, range, and boxplots.

```
summary(boys$bmi); sd(boys$bmi)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.18   19.57   20.90   20.63   21.58   22.46 
```

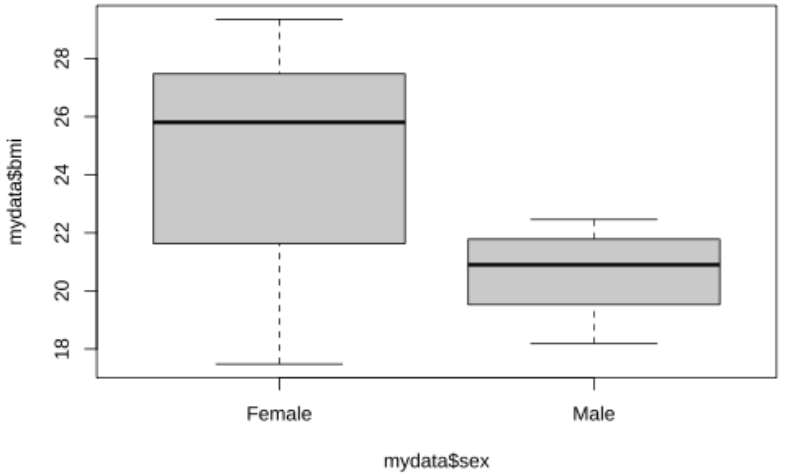
```
[1] 1.466896
```

```
summary(girls$bmi); sd(girls$bmi)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  17.48   21.95   25.80   24.59   27.47   29.35 
```

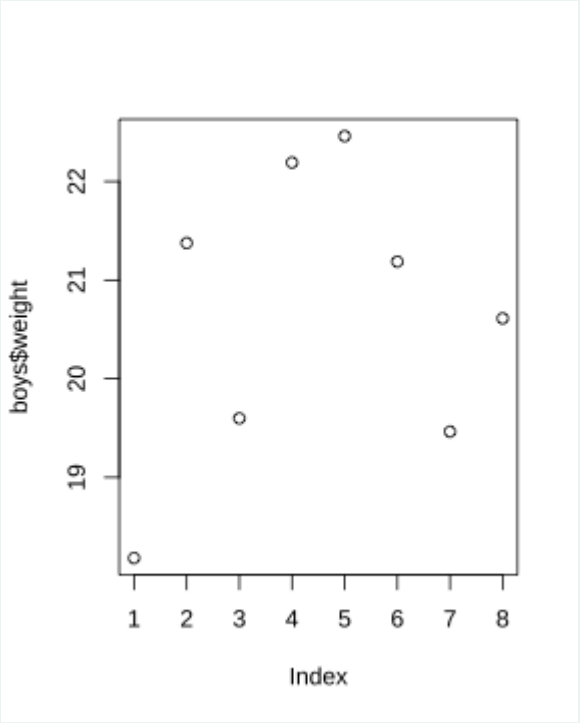
```
[1] 3.70739
```

```
boxplot(mydata$bmi ~ mydata$sex)
```

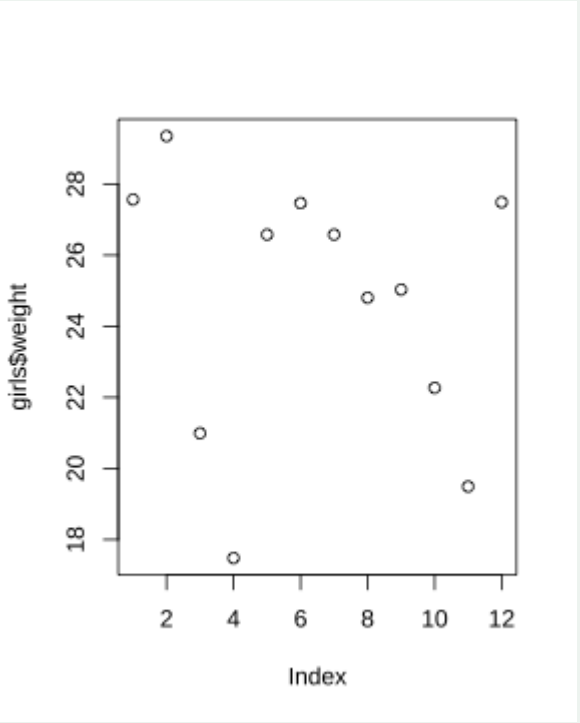


#4 Plot BMI vs. weight for each gender separately. Do they have similar relationships?

```
plot(boys$bmi, boys$weight)
```




```
plot(girls$bmi, girls$weight)
```



#5 Are males or females more likely to be bullied in the past 12 months? Calculate the percentage bullied for each gender.

```
bullied_boys <-
  boys[boys$bullied_past_12mo == TRUE,]
nrow(bullied_boys)
```

```
[1] 3
```

```
bullied_boys_prct <-
  nrow(bullied_boys) / nrow(boys) * 100
bullied_boys_prct
```

```
[1] 37.5
```

```
# alternative
mean(boys$bullied_past_12mo, na.rm=TRUE)
```

```
[1] 0.375
```

```
# Apply the same method for girls:
bullied_girls <-
  girls[girls$bullied_past_12mo == TRUE,]
nrow(bullied_girls)
```

```
[1] 6
```

```
bullied_girls_prct <-
  nrow(bullied_girls) / nrow(girls) * 100
bullied_girls_prct
```

```
[1] 50
```

```
# alternative. Answers don't match. Why???
mean(girls$bullied_past_12mo, na.rm=TRUE)
```

```
[1] 0.4
```

#5 cont'd

On the previous slide we saw that our two methods for calculating the percentage of girls that were bullied in the past 12 months did not match. What went wrong?

```
nrow(bullied_girls)
```

```
[1] 6
```

```
girls$bullied_past_12mo
```

```
 [1]    NA    NA  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE
[11] FALSE FALSE
```

To get the number of girls that were bullied we need to make sure the missing values (NA) are not included.

#5 cont'd - working with NA's

```
# values of bullied_past_12mo
girls$bullied_past_12mo
```

```
 [1]    NA    NA  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE
[11] FALSE FALSE
```

```
# which are missing (logical)
is.na(girls$bullied_past_12mo)
```

```
 [1]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[11] FALSE FALSE
```

```
# which are NOT missing (logical)
!is.na(girls$bullied_past_12mo)
```

```
 [1] FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[11]  TRUE  TRUE
```

#5 cont'd - fix girls' code

Exclude the missing values from the `bullied_girls`:

```
girls2 <- girls[!is.na(girls$bullied_past_12mo),]
nrow(girls2)
```

```
[1] 10
```

```
bullied_girls2 <- girls2[girls2$bullied_past_12mo == TRUE,]
nrow(bullied_girls2)
```

```
[1] 4
```

```
# from girls dataset, total number bullied
sum(girls$bullied_past_12mo, na.rm = TRUE)
```

```
[1] 4
```

#5 cont'd - Calculate percentage girls bullied

```
bullied_girls_prct2 <- nrow(bullied_girls2) / nrow(girls2) * 100
bullied_girls_prct2
```

```
[1] 40
```

```
# Compare to alternative
mean(girls$bullied_past_12mo, na.rm=TRUE)
```

```
[1] 0.4
```

Practice 5

1. Continue adding code chunks to your Rmd (or, start a new one! But remember to load the libraries and data at the top.)

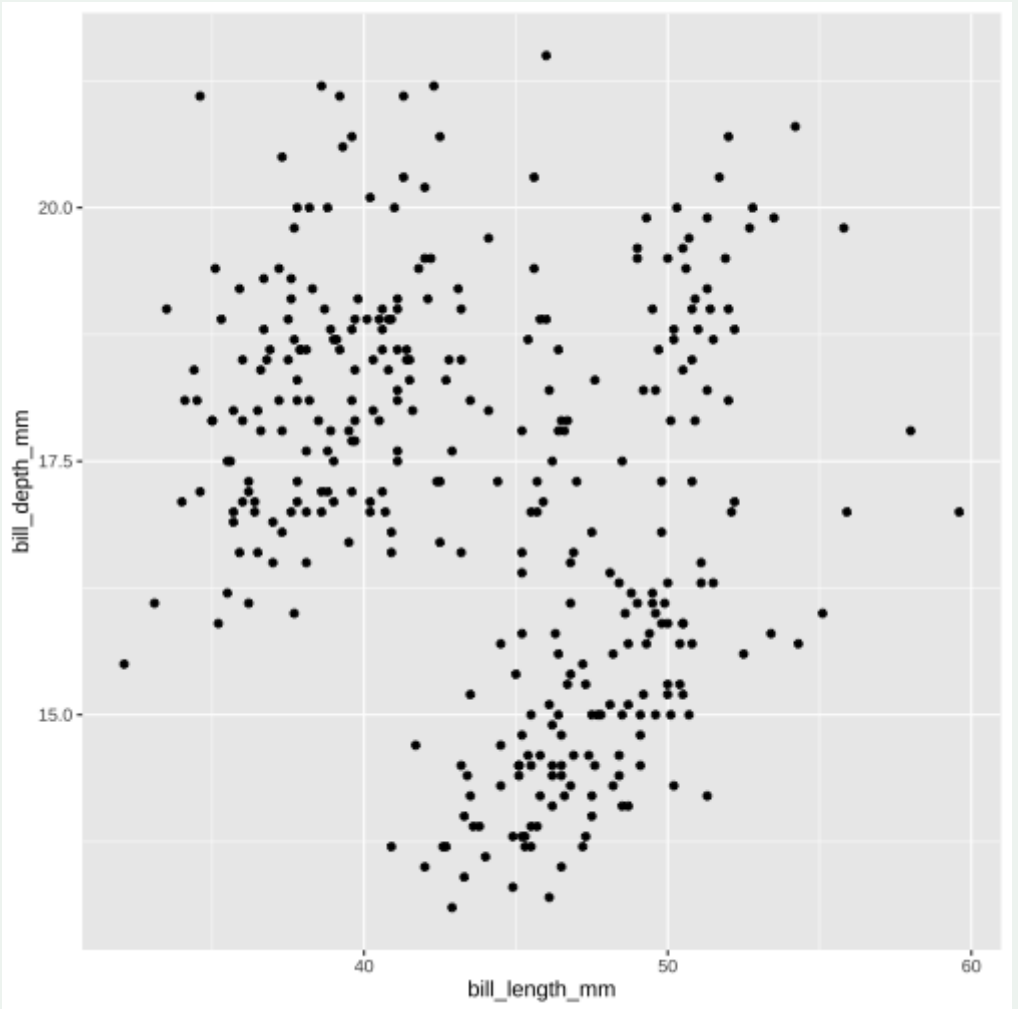
2. Make a scatter plot of bill depth vs bill length.

3. Add `+ geom_smooth(method="lm")` to the plot. What is this saying about the association between bill depth and length?

4. Now add `color = species` to the aesthetic `aes()`. Keep `geom_smooth`. How do the associations look now?

2. Make a scatter plot of bill depth vs bill length.

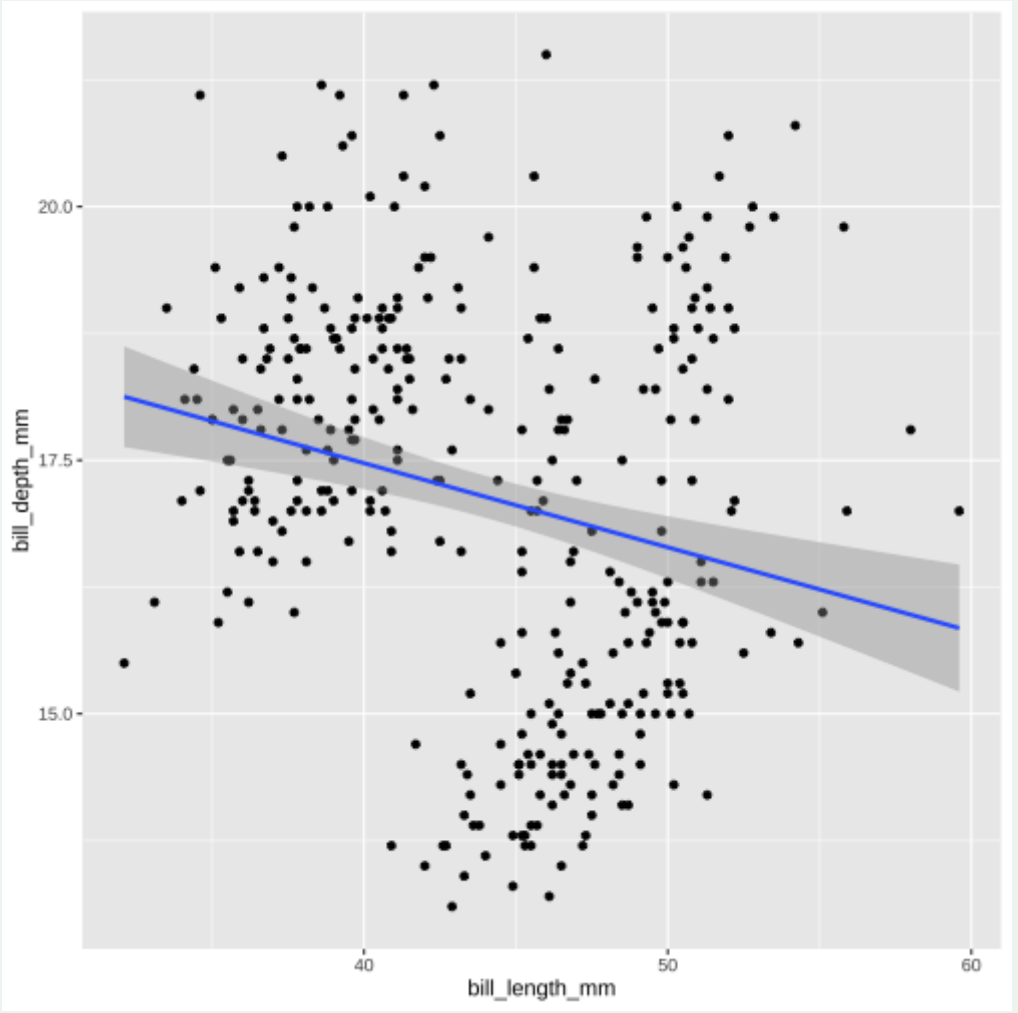
```
ggplot(data = penguins,
       aes(x = bill_length_mm,
           y = bill_depth_mm)) +
  geom_point()
```



3. Add + geom_smooth(method="lm") to the plot. What is this saying about the association between bill depth and length?

It looks like as bill length increases, bill depth decreases. This is a negative association. But there is a wide variation and a lot of noise.

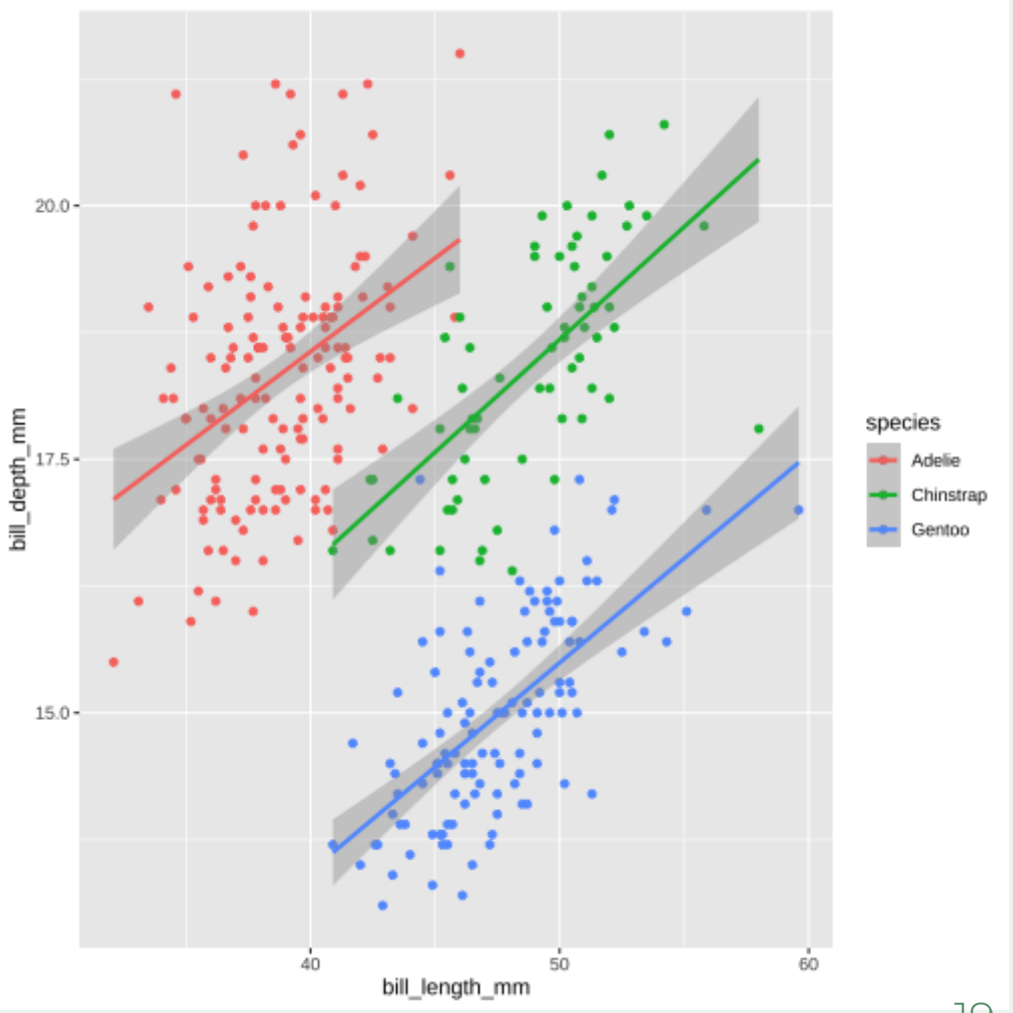
```
ggplot(data = penguins,
       aes(x = bill_length_mm,
           y = bill_depth_mm)) +
  geom_point()+
  geom_smooth(method="lm")
```



4. Now add `color = species` to the aesthetic `aes()`. Keep `geom_smooth`. How do the associations look now?

The association reverses, when we look inside species. As bill length increases, bill depth increases. This is an example of Simpson's paradox!

```
ggplot(data = penguins,
       aes(x = bill_length_mm,
           y = bill_depth_mm,
           color = species)) +
  geom_point()+
  geom_smooth(method="lm")
```



Here's a prettier version:

```
ggplot(data = penguins,
       aes(x = bill_length_mm,
           y = bill_depth_mm,
           color = species)) +
  geom_point()+
  geom_smooth(method="lm")+
  labs(
    title = "Flipper vs bill length",
    subtitle = "Palmer Station LTER",
    x = "Flipper length(mm)",
    y = "Bill length(mm)") +
  scale_color_viridis_d(
    name = "Species") +
  theme(legend.position = "bottom")+
  theme_bw()
```