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

Create a new Rmd or continue in your current Rmd.

1. Create a dataset for just the Torgersen island penguins that are female.
2. Restrict the data to just Torgersen female penguins that weigh more than 3500 g.
3. Restrict the dataset from the previous step to just the columns with the original body measurements.
4. Add a column for the difference in the flipper and bill lengths, and call it `flipper_bill_diff`.
5. How many rows and columns does your final dataset have?

Practice 4 Answers

#1 Create a dataset for just the Torgersen island penguins that are female.

```
Torg_female <- penguins %>%
  filter(island == "Torgersen" & sex == "female")
```

#2 Restrict the data to just Torgersen female penguins that weigh more than 3500 g.

```
Torg_female2 <- Torg_female %>%
  filter(body_mass_g > 3500)
```

#3 Restrict the dataset from the previous step to include just the columns with the original body measurements.

```
Torg_female3 <- Torg_female2 %>%
  select(bill_length_mm:body_mass_g)
```

#4 Add a column for the difference in the flipper and bill lengths, and call it `flipper_bill_diff`.

```
Torg_female4 <- Torg_female3 %>%
  mutate(flipper_bill_diff = flipper_length_mm - bill_length_mm)
```

#5 How many rows and columns does your final dataset have?

```
dim(Torg_female4)
```

```
[1] 9 5
```

9 rows and 5 columns.

Note: Steps 1-4 could have been done with consecutive pipes:

```
Torg_female5 <- penguins %>%
  filter(island == "Torgersen" & sex == "female" & body_mass_g > 3500) %>%
  select(bill_length_mm:body_mass_g) %>%
  mutate(flipper_bill_diff = flipper_length_mm - bill_length_mm)
dim(Torg_female5)
```

```
[1] 9 5
```

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.)

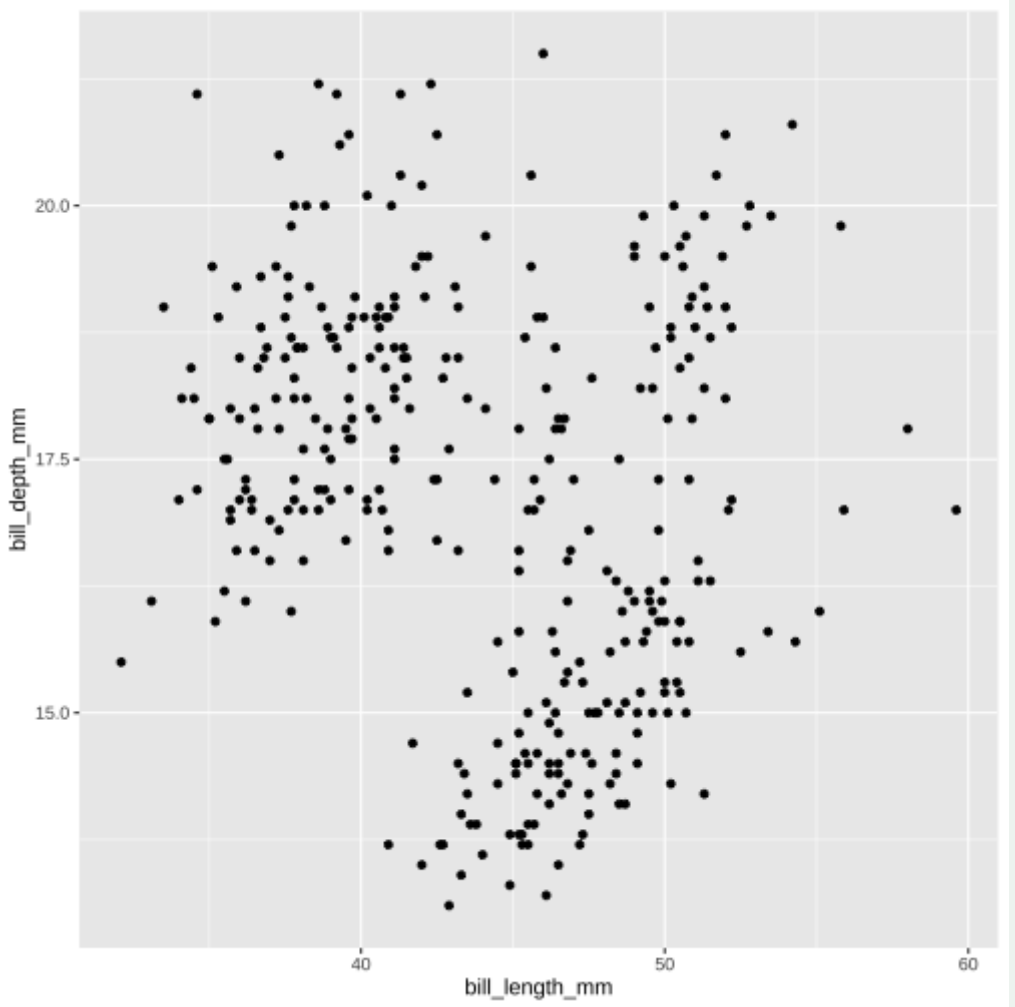
2. Make a scatter plot of bill depth vs bill length.

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.

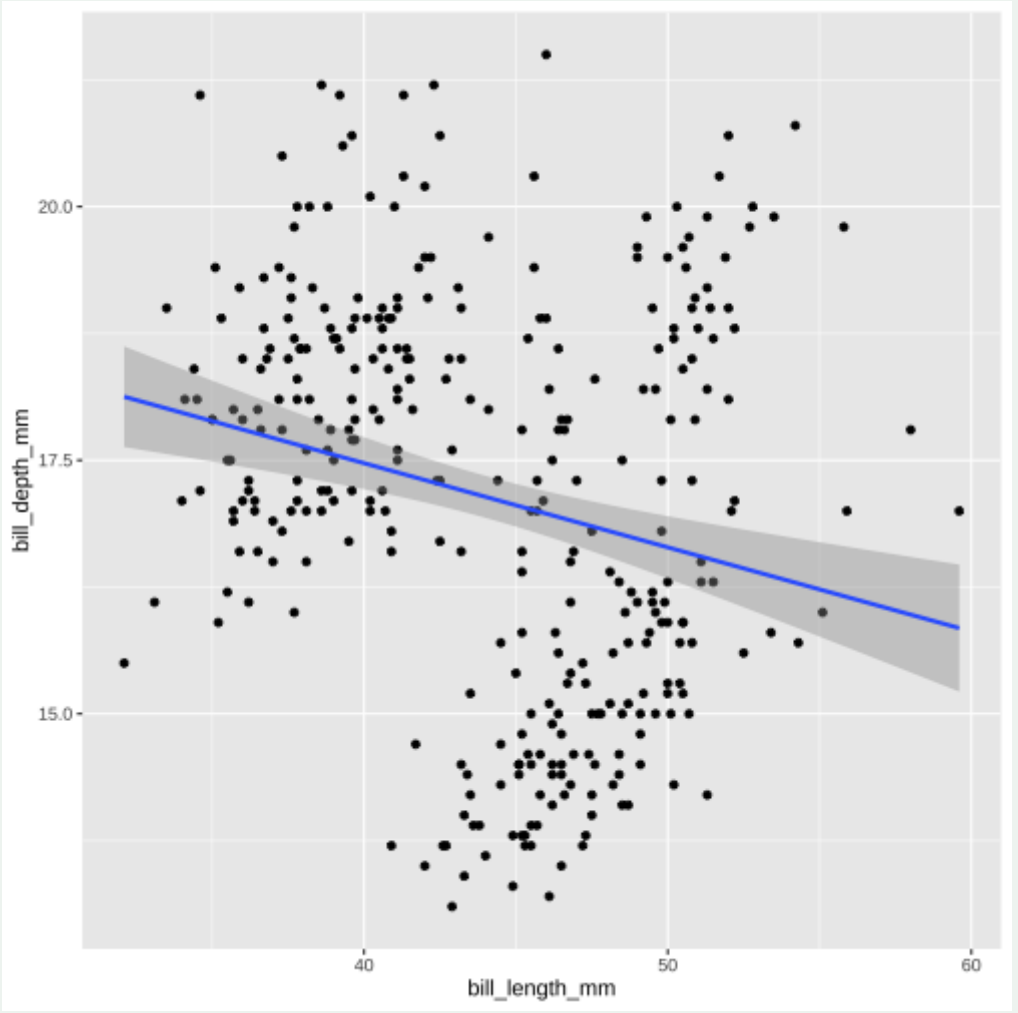
```
ggplot(data = penguins,
       aes(x = bill_length_mm,
           y = bill_depth_mm)) +
  geom_point()
```



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.

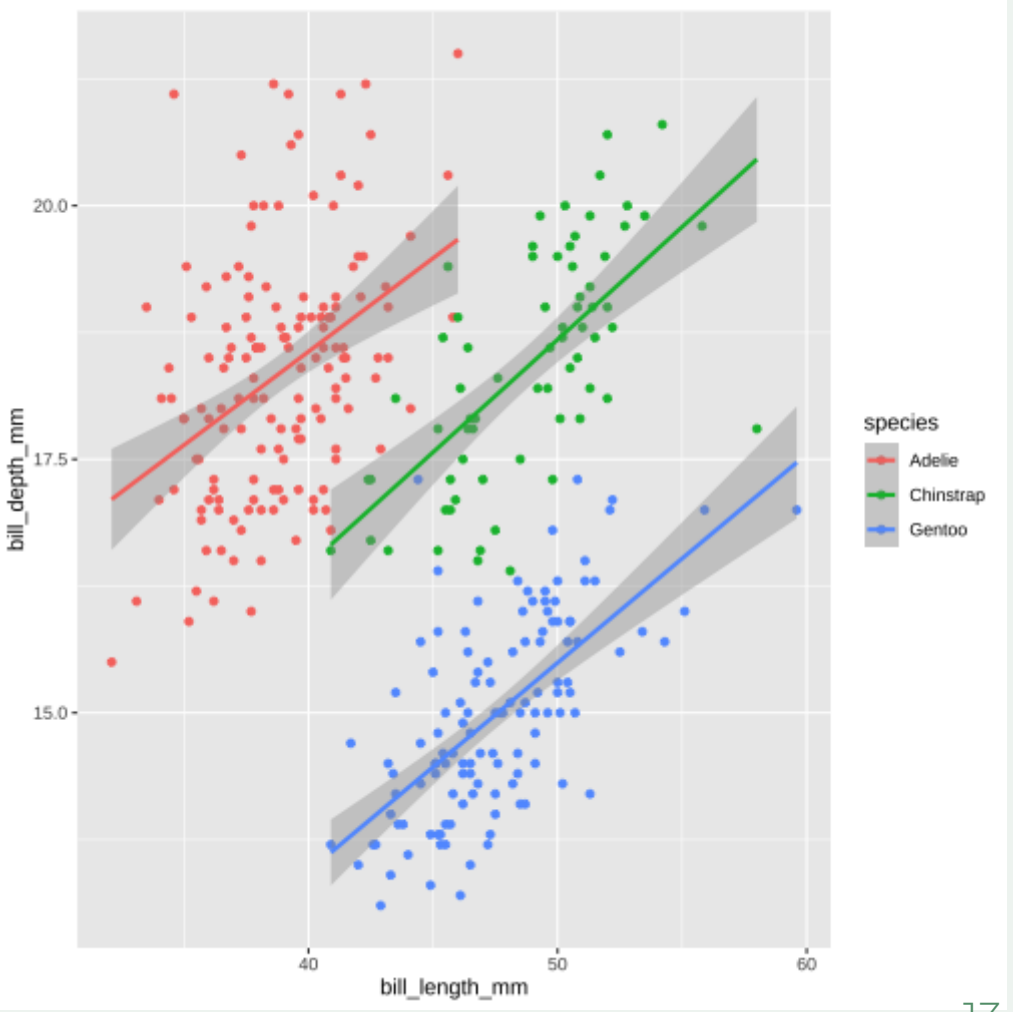
```
ggplot(data = penguins,
       aes(x = bill_length_mm,
           y = bill_depth_mm)) +
  geom_point()+
  geom_smooth(method="lm")
```



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()
```