S&DS 220: Homework 8

Due Friday April 5

Braeden Cullen

Instructions

1. Complete the questions below. Upload your knitted PDF solutions to Gradescope by the due date.
2. Your solutions should be a combination of writing and R code. When writing, use complete sentences.
3. Previous homework assignments already had code chunks created for you. Now it is up to you to insert R code chunks within each problem as needed.
4. You should aim for clear and concise communication (in both words and R code).

Problem set questions

Question 1: COVID-19 data

Work through the Vignette: COVID-19 at the end of chapter 7. Read through the vignette, copying and running the code in the book for yourself in a separate R script. *There is nothing you need to turn in for this task in this assignment.*

Question 2: (6.34) Billboard music charts

The exercise uses the `billboard` data from the `tidyr` package.

(a) Which artist had the most tracks on the chart in 2000?

```
tidyr::billboard |>
  group_by(artist) |>
  summarise(n = n()) |>
  slice_max(n, n=1) |>
  arrange(desc(n)) |>
  head(1)
```

```
## # A tibble: 1 x 2
##   artist     n
##   <chr>  <int>
## 1 Jay-Z      5
```

(b) Which track from 2000 spent the most weeks at #1?

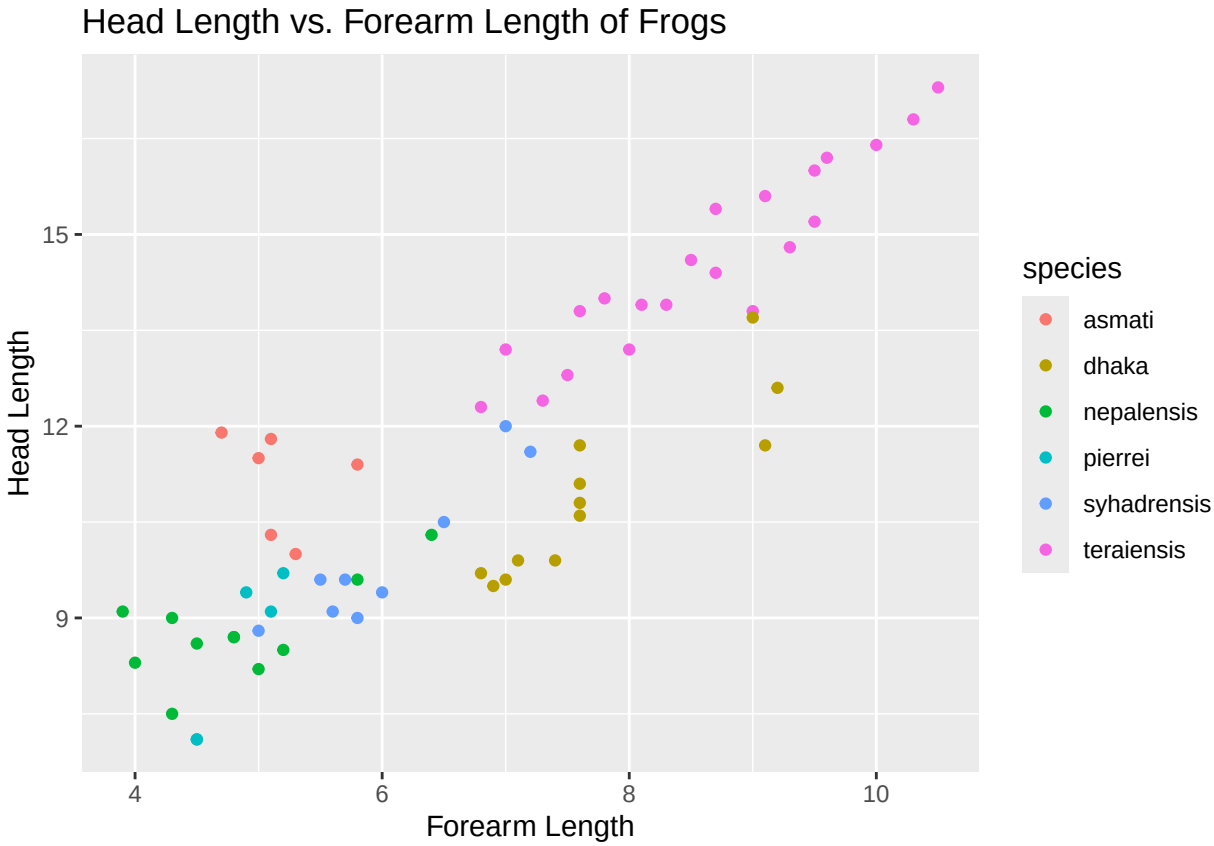
```
tidyr::billboard |>
  pivot_longer(starts_with("wk"), names_to = "week", values_to = "rank") |>
  filter(rank == 1) |>
  group_by(track) |>
  summarise(n = n()) |>
  slice_max(n, n=1) |>
  arrange(desc(n)) |>
  head(1)
```

```
## # A tibble: 1 x 2
##   track                       n
##   <chr>                   <int>
## 1 Independent Women Pa...    11
```

Question 3: (7.27) A new frog?

Consider the `frogs` data set in the `fosdata` package. This data was used to argue that a new species of frog had been found in a densely populated area of Bangladesh. Create a scatterplot of head length distance from tip of snout to back of mandible versus forearm length distance from corner of elbow to proximal end of outer palmar metacarpal tubercle, colored by species. Explain whether this plot is visual evidence that the physical characteristics of the dhaka frog are different than the other frogs. Give the plot an appropriate title and axes names.

```
fosdata::frogs %>%
  ggplot(aes(x = fal, y = hl, col = species)) +
  geom_point() +
  labs(title = "Head Length vs. Forearm Length of Frogs",
       x = "Forearm Length",
       y = "Head Length")
```



Explain whether this plot is visual evidence that the physical characteristics of the dhaka frog are different than the other frogs:

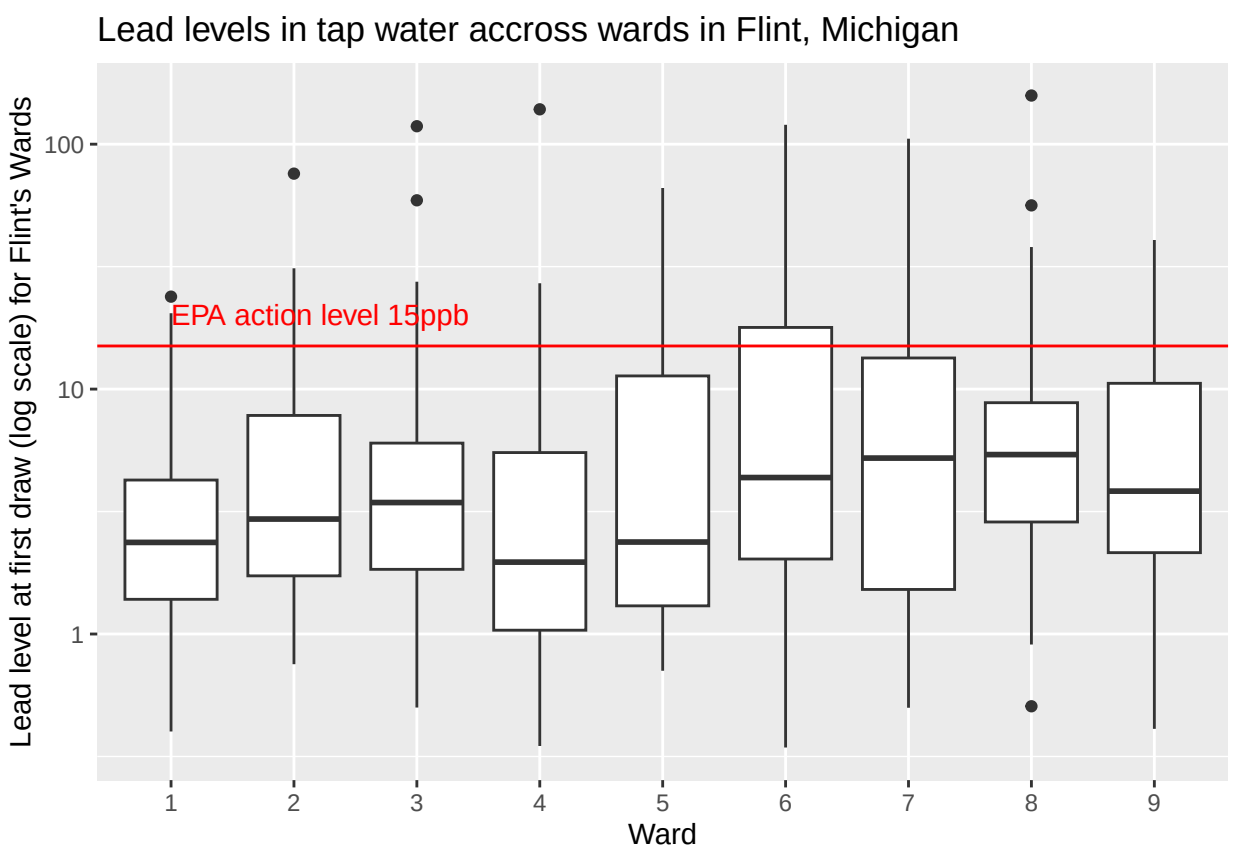
This plot is visual evidence that the physical charictaristics of the dhaka frog are different than the other frogs. The dhaka frog has a large forearm length and a comparatively small head length compared to the other frogs. The other frogs with similarly long forearm lengths also have longer head lengths, while the dhaka frog has a much shorter head length. This trend is consistent accross all the other frogs, with the exception of the dhaka frog thus providing evidence that the physical characteristics of the dhaka frog are different than the other frogs.

Question 4: (7.32) Recreating a polished plot of water data

The data `flint` from `fosdata` gives the results of tap water lead testing during the Flint, Michigan water crisis in 2015. Figure 7.25 in the text is a graph showing lead levels at first draw (Pb1) for Flint's eight geographical areas, called "Wards". The red horizontal line represents the EPA's "action level" for lead in water, at 15 ppb. Reproduce this graph as well as you can (see the text).

The y-axis scale is logarithmic, which you can accomplish with `scale_y_log10()`. Note that there is no Ward 0 in Flint.

```
fosdata::flint %>%
  filter(Ward != 0) %>% # must remove ward 0
  ggplot(aes(x = Ward, y = Pb1)) +
  geom_boxplot() +
  geom_hline(yintercept = 15, col = "red") +
  scale_y_log10() +
  geom_text(inherit.aes = FALSE,
            data = tibble(x = 1, y = 15, label = "EPA action level 15ppb"),
            aes(x = x, y = y, label = label),
            hjust = 0, vjust = -1, col = "red") +
  labs(title = "Lead levels in tap water accross wards in Flint, Michigan",
       x = "Ward",
       y = "Lead level at first draw (log scale) for Flint's Wards")
```



Question 5: (7.26) *Twister* movie ratings

This question uses the movies data set `movies_wide.rds` in the "Files -> Data sets" directory in Canvas

Create a scatterplot of the ratings of Twister (1996) versus the date of review, and add a trend line using `geom_smooth`. Give the plot an appropriate title and axes names.

```
fosdata::movies %>%
  filter(title == "Twister (1996)") %>%
  ggplot(aes(x = lubridate::as_datetime(timestamp), y = rating)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Ratings of Twister (1996) Over Time",
       x = "Date of Review",
       y = "Rating")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```