Answer Sketch for Homework E

431 Staff and Professor Love

*Due **2019-09-27** at 2 PM. Last Edited 2019-09-27 15:18:27*

Contents

R Setup

Here's the complete R setup we used.

```
knitr::opts_chunk$set(comment=NA)
options(width = 60)

library(janitor); library(magrittr); library(tidyverse)
```

Looking over the `iris1` tibble

We'll start by creating a tibble for the iris data.

```r
iris1 <- tbl_df(iris)
```

Here are the first few rows of the tibble.

```r
iris1
```

```
# A tibble: 150 x 5
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
          <dbl>       <dbl>        <dbl>       <dbl> <fct>  
 1          5.1         3.5          1.4         0.2 setosa 
 2          4.9         3            1.4         0.2 setosa 
 3          4.7         3.2          1.3         0.2 setosa 
 4          4.6         3.1          1.5         0.2 setosa 
 5          5           3.6          1.4         0.2 setosa 
 6          5.4         3.9          1.7         0.4 setosa 
 7          4.6         3.4          1.4         0.3 setosa 
 8          5           3.4          1.5         0.2 setosa 
 9          4.4         2.9          1.4         0.2 setosa 
10          4.9         3.1          1.5         0.1 setosa 
# ... with 140 more rows
```

Incidentally, here's what these species look like.

Question 1

> Across the entire sample of 150 flowers, find and interpret the correlation of petal length and petal width. Does it matter much whether you use the Pearson or Spearman correlation?

The correlation of petal length and petal width is very strong and positive in these data. In this case it doesn't much matter which measure of correlation we use, as the estimates are very similar for these data.

Before we see the picture (which will happen in Question 2, of course), it seems as though there will be a strong linear association with a positive slope (so that the flowers with, for instance, larger petal widths are also generally going to be the flowers with larger petal lengths.)

Table of Correlation Estimates

I'll create a little tibble of results here.

```
q1 <- iris1 %>%
  summarize(r.pearson = cor(Petal.Length, Petal.Width),
      r.spearman = cor(Petal.Length, Petal.Width, method = "spearman")) %>%
  round(digits = 3)

knitr::kable(q1,
        caption = "Correlations of Petal Length and Petal Width")
```

Table 1: Correlations of Petal Length and Petal Width

r.pearson	r.spearman
0.963	0.938

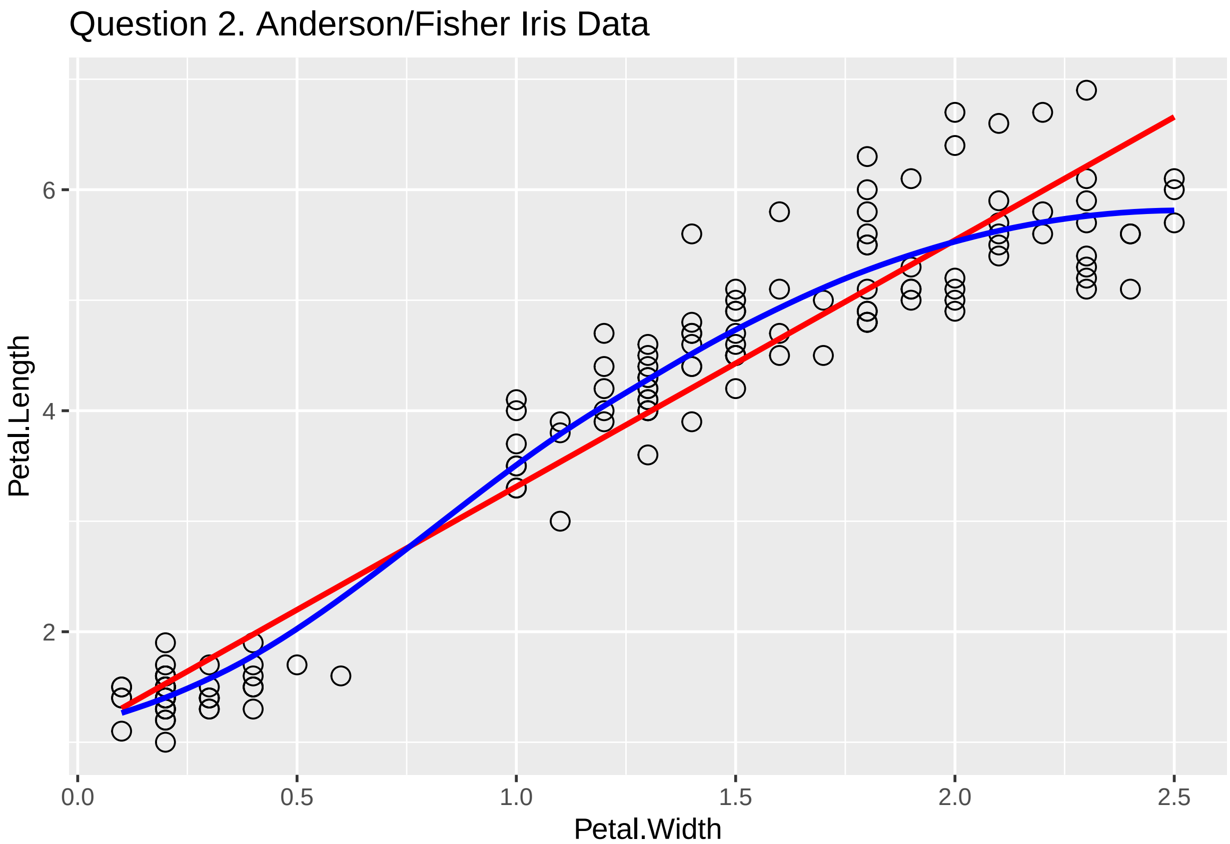
The Pearson correlation, at 0.963, is slightly larger than the Spearman correlation of 0.938 between the petal lengths and widths. In either case, the correlation is very positive and strong.

Question 2

> Draw an appropriate scatterplot to assess the prediction of petal length (our *outcome*) using petal width as a predictor. Include both a loess smooth and regression line in your plot. Does the plot suggest that a straight line model is appropriate in this case?

Here is my scatterplot.

```
ggplot(iris1, aes(x = Petal.Width, y = Petal.Length)) +
    geom_point(size = 3, shape = 1) + ## default size = 2
    geom_smooth(method = "lm", se = FALSE, color = "red") +
    geom_smooth(method = "loess", se = FALSE, color = "blue") +
    labs(title = "Question 2. Anderson/Fisher Iris Data")
```



A straight line model certainly fits most of the data reasonably well, although it looks like there are at least two groups of flowers (the smaller ones, with petal width less than 1 cm, as compared to the larger ones.) The pattern of the smaller flowers is less obviously on the proposed straight line than are the larger flowers, in my view.

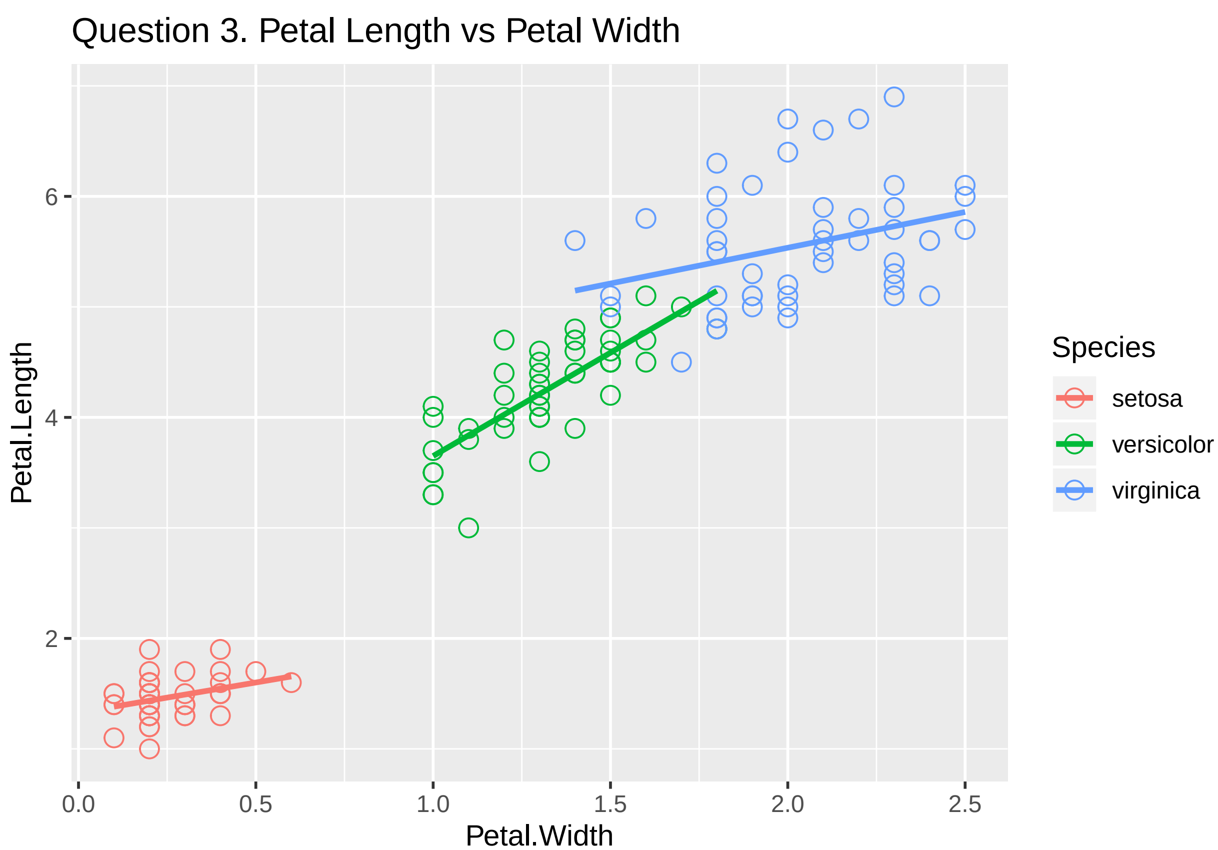
Question 3

> Suppose we are interested in which of the three types of iris (*setosa*, *versicolor* or *virginica*) shows the strongest *linear* relationship between petal length and petal width. Draw an attractive and thoughtfully labeled plot of the relevant data to address this issue, accompanied by a sentence or two describing the key findings from your plot. Postpone the discussion of a numerical summary to Question 4.

All three plots show at least reasonable adherence to a straight line model, although none of them appear to be as strong as the association we saw in the complete data. In my opinion, the *versicolor* model looks most promising, and *versicolor* also clearly has the most pronounced positive slope.

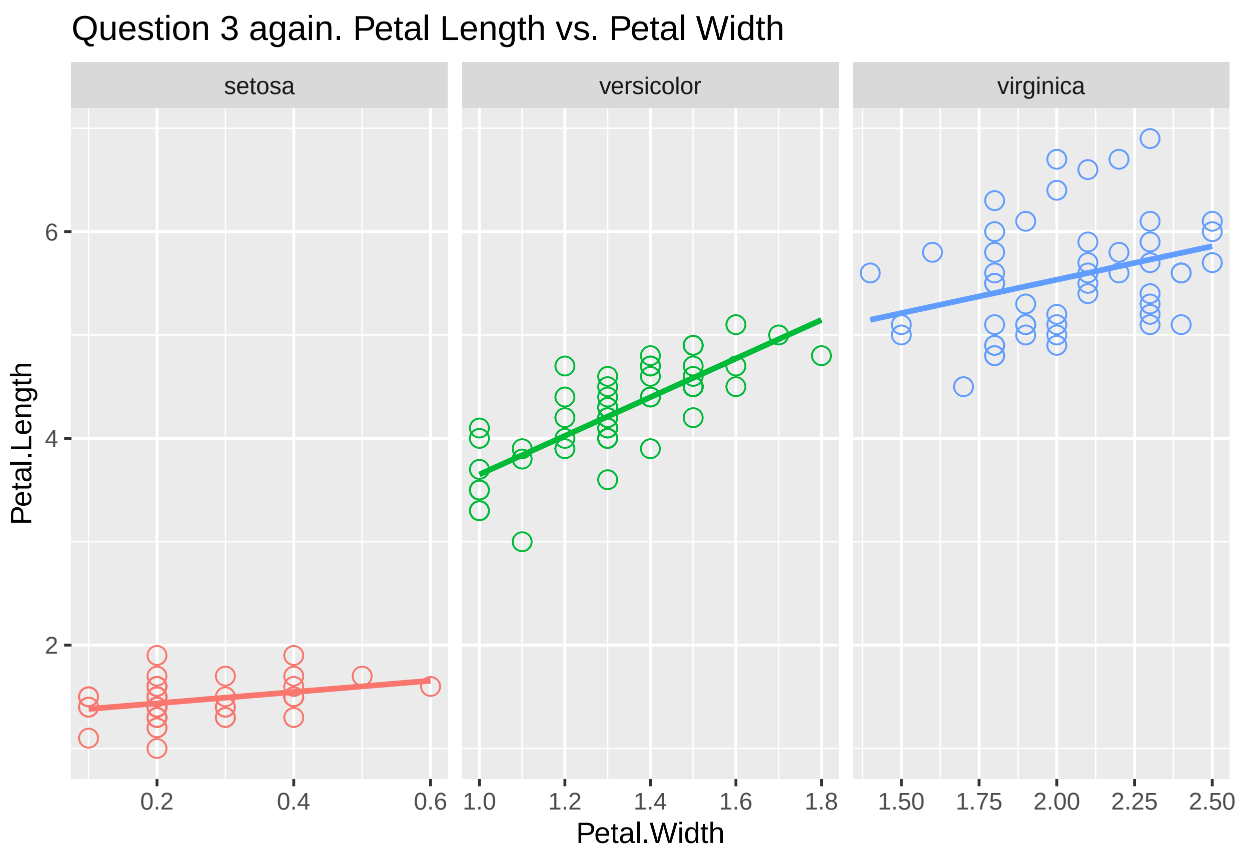
I started with this plot.

```
ggplot(iris1, aes(x = Petal.Width, y = Petal.Length, color = Species)) +
  geom_point(size = 3, shape = 1) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Question 3. Petal Length vs Petal Width")
```



As an alternative, the plot below permits the Petal Length tick marks on the horizontal axis to vary freely across the three facetted scatterplots.

```
ggplot(iris1, aes(x = Petal.Width, y = Petal.Length, color = Species)) +
  geom_point(size = 3, shape = 1) +
  geom_smooth(method = "lm", se = FALSE, fullrange = TRUE, aes(color = Species)) +
  labs(title = "Question 3 again. Petal Length vs. Petal Width") +
  facet_wrap(~ Species, scales = "free_x") +
  guides(color = FALSE)
```



Some people tried to fit a loess smooth to the data, which seems to fall apart. That's because of the minimal variation in Petal Wicths among the *setosa*.

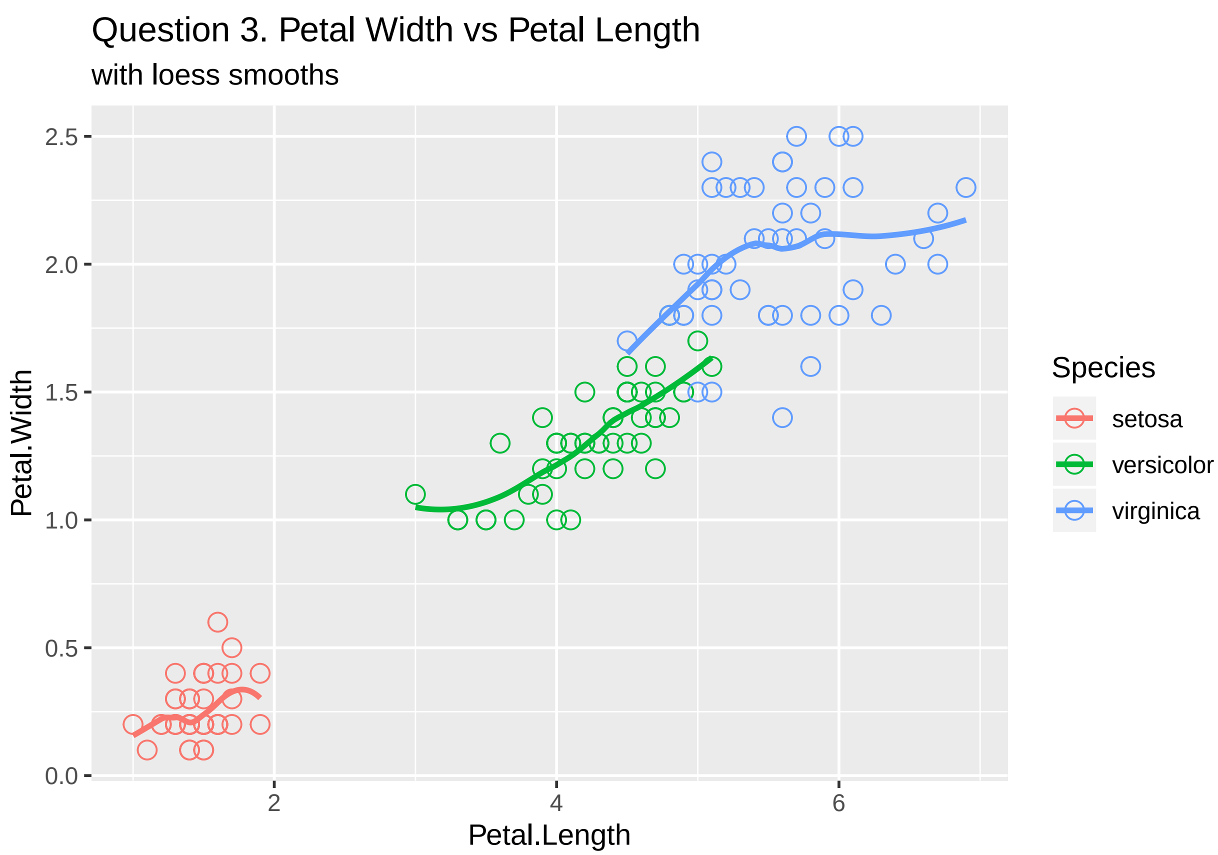
```
iris1 %>% filter(Species == "setosa") %>%
  tabyl(Petal.Width)
```

```
 Petal.Width  n percent
         0.1  5    0.10
         0.2 29    0.58
         0.3  7    0.14
         0.4  7    0.14
         0.5  1    0.02
         0.6  1    0.02
```

See how there are only six values of Petal Width observed, and the vast majority of them are at a single point (0.2) while two values (0.5 and 0.6) have only one observation. The loess smooth collapses in this case, as it really is designed to address situations where we have meaningful variation in the X *and* Y variables, and can thus borrow strength to improve predictions from nearby observations with differing X values. Here, there just aren't enough data to support a complicated model like a loess smooth.

Another option, I suppose, was for Question 3 to look instead at the relationship with Length on the X axis and Width on the Y. That would at least alleviate the big problem, although it no longer matches the setup from the other questions in this Homework so well.

```
ggplot(iris1, aes(x = Petal.Length, y = Petal.Width, color = Species)) +
  geom_point(size = 3, shape = 1) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Question 3. Petal Width vs Petal Length",
       subtitle = "with loess smooths")
```



Question 4

> Is the *correlation* between petal length and petal width larger or smaller in the sample of 150 flowers than in the species-specific samples of 50 flowers each? Why does this appear to be the case? On the basis of your correlation, which iris type shows the strongest *linear* relationship for the prediction of petal length on the basis of petal width? Specify the Pearson correlation (to two decimal places) for your strongest model here.

The correlation between petal length and petal width for the entire sample of 150 flowers is 0.96, and this turns out to be far larger than any of the species-specific correlations.

Here are the Pearson correlation coefficients for the three species, sorted in descending order of their values:

```
petal_corr <- iris1 %>%
  group_by(Species) %>%
```

```
  summarize(Correlation = round(cor(Petal.Length, Petal.Width), digits = 2)) %>%
  arrange(desc(Correlation))

knitr::kable(petal_corr,
        caption = "Pearson correlations for Petal Length vs Width")
```

Table 2: Pearson correlations for Petal Length vs Width

Species	Correlation
versicolor	0.79
setosa	0.33
virginica	0.32

Why does this appear to be the case?

Within each of these three species, the linear relationship is weaker (both in terms of displaying more scatter around the regression line, and the line itself having less steep positive slope) as compared to across the three species. For *versicolor*, these effects are somewhat less pronounced so that the linear relationship still looks pretty strong, as compared to *setosa* or *virginica*. Iris *versicolor* shows the strongest linear relationship, with a Pearson correlation coefficient of 0.79.

The three species are largely identifiable by petal width - the *setosa* are the smallest group, and the *versicolor* are generally smaller than *virginica*. The species-specific variation isn't strongly linear, but across species, we pick up substantial additional correlation, in that the petal lengths are similarly separated by species.

If you're interested, here's a table of the results for both the Pearson and Spearman correlations, although I think Pearson is more relevant since we're looking specifically for *linear* associations that can be modeled, not merely monotone ones.

```
iris1 %>%
  group_by(Species) %>%
  summarise(Pearson.Corr =
              round(cor(Petal.Length, Petal.Width, method = "pearson"),
                    digits = 2),
            Spearman.Corr =
              round(cor(Petal.Length, Petal.Width, method = "spearman"),
                    digits = 2)) %>%
  arrange(desc(Pearson.Corr)) %>%
  knitr::kable(caption = "Petal Length vs. Width Correlations")
```

Table 3: Petal Length vs. Width Correlations

Species	Pearson.Corr	Spearman.Corr
versicolor	0.79	0.79
setosa	0.33	0.27
virginica	0.32	0.36

Question 5

> Using the strongest model that you identified in question 4, what is the difference in the predicted petal length between two new flowers, one of whom has a petal width at the 75th percentile of the original data for that iris type, and the other of whom has a petal width at the 25th percentile of the original data for that iris type? Be sure to specify which of the two new flowers would be expected to be longer.

The linear model for the *iris versicolor* data is as follows. First, we'll filter to create a new tibble with just the versicolor data:

```
iris1.ver <- iris1 %>% filter(Species == "versicolor")
```

Then, we'll make the linear model:

```
ver.petal.lm <- lm(Petal.Length ~ Petal.Width, data = iris1.ver)
ver.petal.lm
```

```

Call:
lm(formula = Petal.Length ~ Petal.Width, data = iris1.ver)

Coefficients:
(Intercept)  Petal.Width  
      1.781        1.869  
```

We'll also need the quartiles for the Petal Widths in the *iris versicolor* data:

```
mosaic::favstats(~ Petal.Width, data = iris1.ver)
```

```
Registered S3 method overwritten by 'mosaic':
  method                           from   
  fortify.SpatialPolygonsDataFrame ggplot2

 min  Q1 median  Q3 max  mean        sd  n missing
   1 1.2    1.3 1.5 1.8 1.326 0.1977527 50       0
```

So, the predicted petal length for an *iris versicolor* blossom at the 25th percentile (1.2 cm) of petal width is: $1.78 + 1.87\ (1.2) =$ **4.02 cm**

The predicted petal length for an *iris versicolor* blossom at the 75th percentile (1.5 cm) of petal width is: $1.78 + 1.87\ (1.5) =$ **4.59 cm**.

So the difference between those two predicted petal lengths is 4.59 - 4.02 = **0.57 cm.** The larger blossom in terms of petal width is thus predicted to have a larger petal length, as well.

Question 6

> Build an attractive and thoughtfully labeled plot (which might include multiple facets, for instance) of the relationship between *sepal length* and *sepal width*, so that the plot distinguishes between the three types of iris. For example, you might use color to indicate each iris type, and show color-coded loess smooths (or linear fits, your choice) for each iris type. Or be more creative. Describe the conclusions of your plot in a nice caption.

The strong correlation overall between petal length and petal width that we saw earlier is not repeated here in the same way. In particular, we will see from either plot that the *iris setosa* plants have the *smallest* sepal lengths on average, but the *largest* sepal widths.

The two plots that we built are distinguished by the decision to not use (Figure 6a) or to use (Figure 6b) facets to separate out the three species of iris.

Plot without Facets (Figure 6a)

```
ggplot(iris, aes(Sepal.Length, Sepal.Width, color = Species)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Figure 6a. Sepal Width increases
       with Sepal Length, for each Iris species",
       subtitle = "with loess smooths",
       x = "Sepal Length (cm)", y = "Sepal Width (cm)")
```

Plot with Facets (Figure 6b)

```
ggplot(iris, aes(Sepal.Length, Sepal.Width, color = Species)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~ Species) +
  guides(color = FALSE) +
  labs(title = "Figure 6b. Sepal Width increases
       with Sepal Length, for each Iris species",
       x = "Sepal Length (cm)", y = "Sepal Width (cm)")
```

Choosing between Figures 6a and 6b is largely a matter of personal preference, in my view.

Question 7 - Fox or Hedgehog?

We don't write answer sketches for essay questions. We're looking for a clear, coherent piece of writing, written in complete English sentences, that describes the patterns you see in the field of interest you specify. We also want to see an example that is meaningful, clearly presented in your essay, and some thoughtful writing about it. We'll gather a few of the more interesting and enlightening responses, and share de-identified excerpts with the group after grading.

On Grading Homework E

Your grade on Homework E is on a 0-100 scale.

General/Administrative (15 points)

- Award up to 10 points for turning the assignment on time (on time = within 1 hour of the deadline)
  - 10 points for both Markdown and HTML in on time.
  - 6 points for one of Markdown, HTML in on time.
  - 0 points if neither is in on time.
  - If a student hasn't submitted either the Markdown or HTML piece, please identify and pester them via email until they do.
- Award an additional 5 points if there is an on-time answer provided for each of the questions asked, and 0 if not.
- Award zero points on the entire assignment to anyone whose first submission of the assignment is more than 4 hours late, unless excused from the assignment by Professor Love.

Question 1. (5 points)

The student should note that the Pearson correlation (0.96) is larger than Spearman (0.94), and that the correlation appears strongly positive.

- No more than 2/5 points for only providing the numbers for the correlation and not giving any interpretation.

Question 2 (10 points)

We want to see a plot showing the loess smooth and regression line.

- Subtract 3 points if no lines are shown on the plot
- Subtract 5 points if no interpretation of the plot is given (they must answer the following question: Does the plot suggest that a straight line model is appropriate in this case?)

Question 3 (10 points)

Must show at least one plot, for the Iris type that has the strongest correlation, which is Iris versicolor.

- Subtract 2 points for not correctly identifying Iris type with the strongest correlation (Iris versicolor)
- Subtract 2 points for no plot
- Subtract 2 points for no interpretation/description

Question 4 (10 points)

Must answer that correlation is larger in full set of 150 flowers.

- Subtract 5 points if no interpretation/discussion provided
- Subtract 4 points if question is answered fully, but incorrectly

Question 5 (10 points)

- 2 points if they identify the correct type of iris for analysis
- 3 points more if they make the correct linear model after making the subset
- 2 points more for correct 25th percentile and 75th percentile predictions
- 3 points more for presenting the answer appropriately, using complete sentences

Question 6 (10 points)

- up to 6 points for building the plot in an appropriate and attractive way
- up to 4 more for interpreting the plot appropriately, using complete sentences.

Question 7 (30 points)

You need to identify (as a group) the 6-8 best essays (of the complete set of 60) that were read by the TAs (so that's choosing from the best two that each of you read, probably). In the Comments to Professor Love, please briefly identify the top 6-8 and specify the topic of these 6-8 best essays so I can read through them before returning them to the students, and select 2-4 to share.

24 out of 30 points should be given if:

- the essay is clear,
- answers the questions posed,
- meets the word limit, and
- has generally good grammar and spelling.

Students should receive **28-30 points** if they meet all of the standards above, and are one of the 6-8 best essays in the group.

Students should receive **24-27 points** if they meet all of the standards above, but were not in that top group.

Students should receive **20-23 points** if they meet most of the standards above but not all of them.

Students should receive **fewer than 20 points** if they fail to meet at least two of the standards above.

All students should receive some feedback (at least a "Nice job! I found this interesting and well-written") from the TA who did the initial grading of the work. I expect somewhere between 40 and 50 of our 60 students will receive grades on the essay between 20 and 27.