Answer Sketch for Homework 4

431 Staff and Professor Love

'Due 2018-09-28, version 2018-09-16

Contents

0.1 R Setup

Here's the complete R setup we used.

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

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

0.2 Incidentally, here's what these species look like.

0.3 Looking over the iris1 tibble

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

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

Here are the first few rows of the tibble.

```
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
```

1 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.)

1.1 Table of Correlation Estimates

I'll create a little tibble of results here.

```r
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

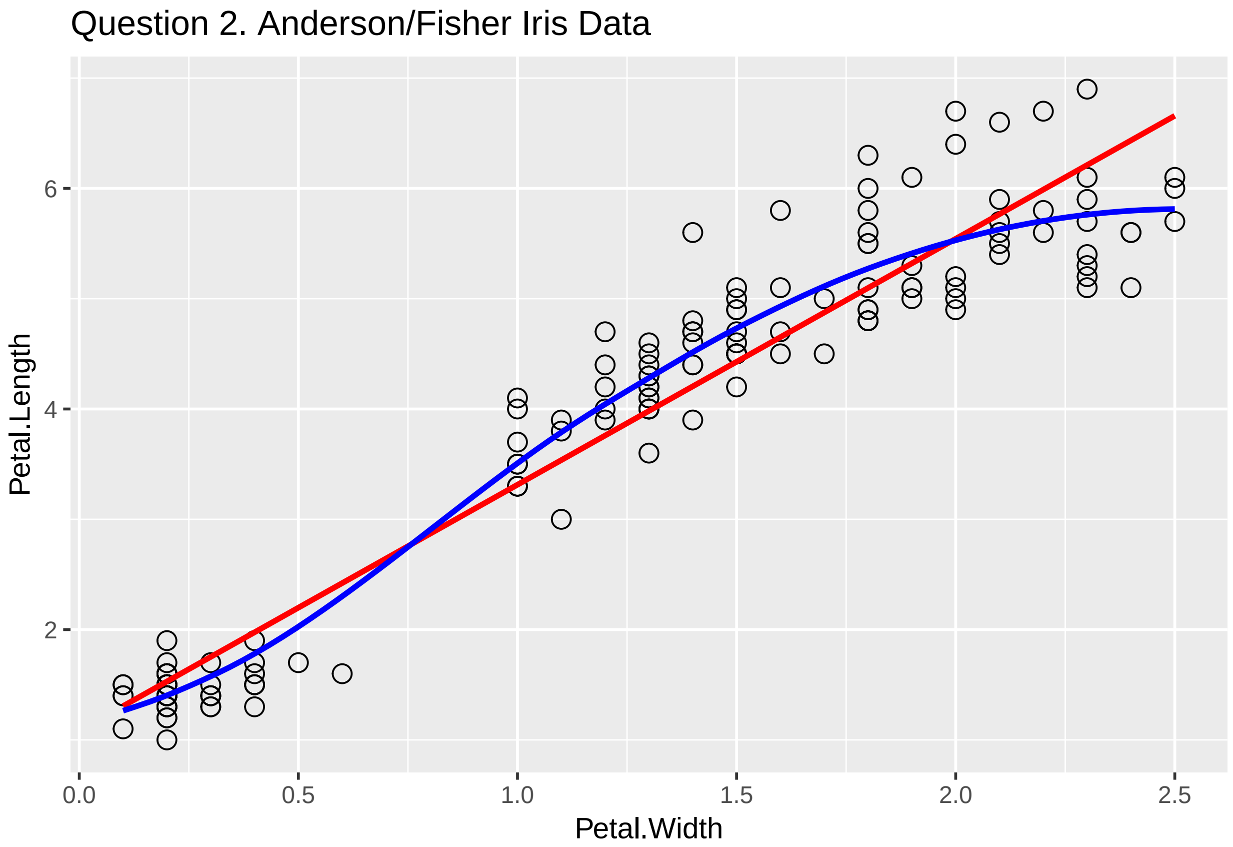
The Pearson correlation, at 0.963, is slightly larger than the Spearman correlation of 0.938 between the petal lengths and widths. In either case, the correlation is very positive and strong.

2 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.

```r
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.

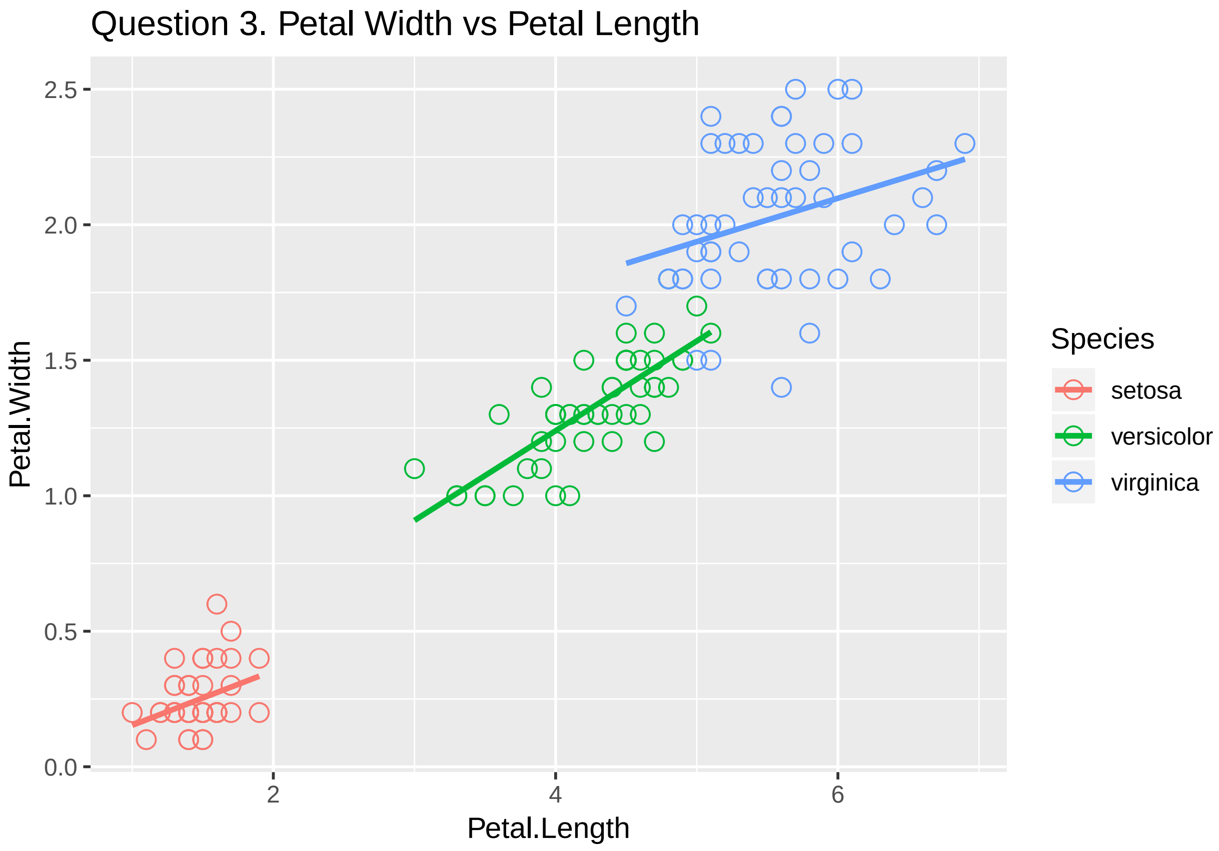
3 Question 3

> 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.

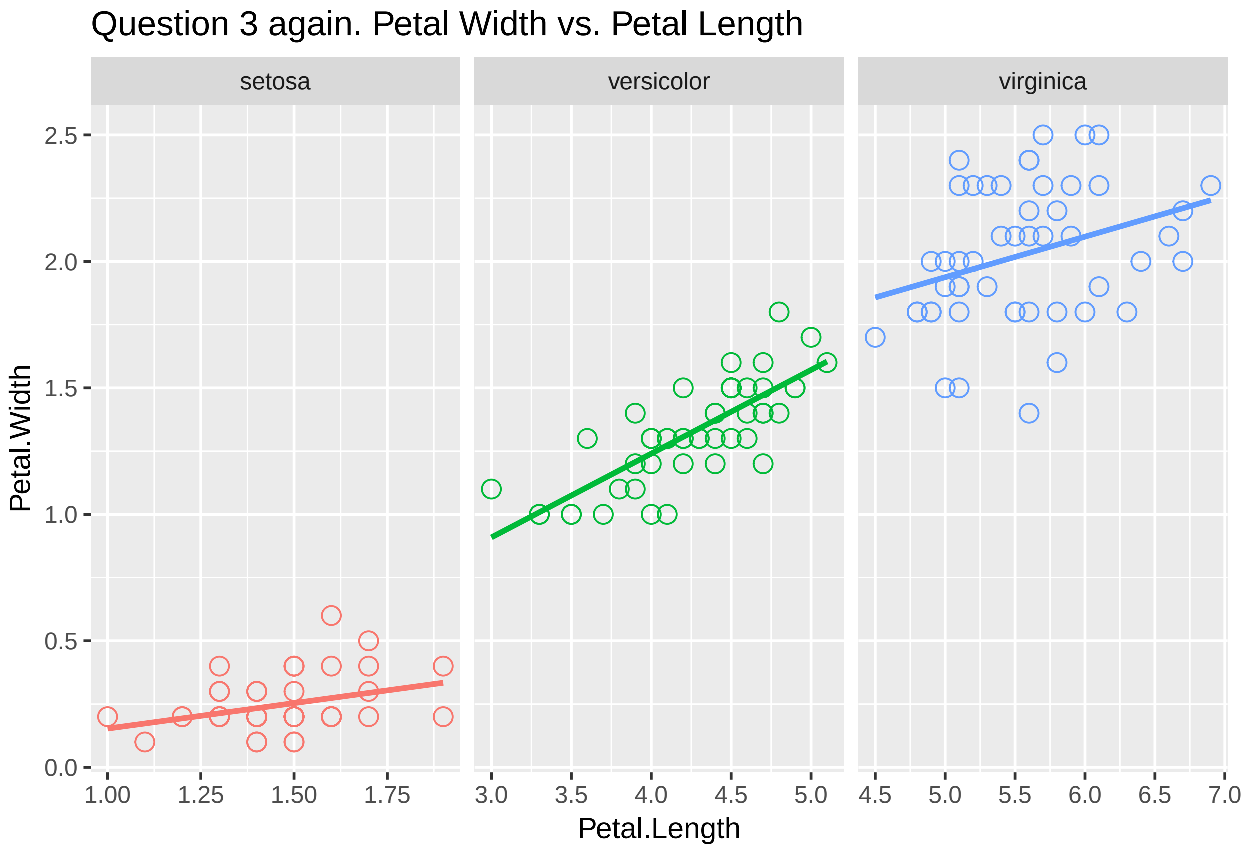
I started with this plot.

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



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.Length, y = Petal.Width, 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 Width vs. Petal Length") +
  facet_wrap(~ Species, scales = "free_x") +
  guides(color = FALSE)
```



4 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

4.1 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

5 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)
```

```
 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.

6 Question 6

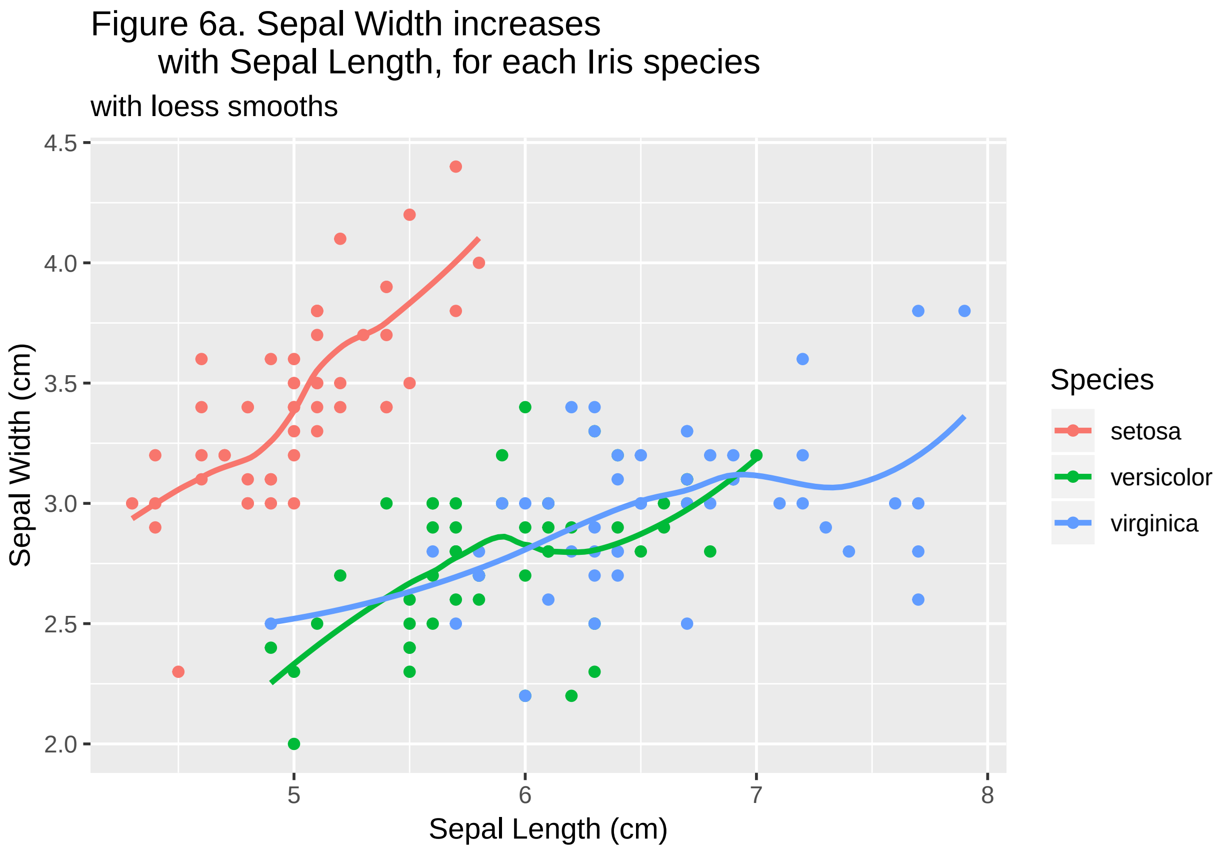
> Build an attractive and thoughtfully labeled plot (which might include multiple facets, for instance) of the relationship between *sepal length* and *sepal width*, so that the plot distinguishes between the three types of iris. For example, you might use color to indicate each iris type, and show color-coded 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.

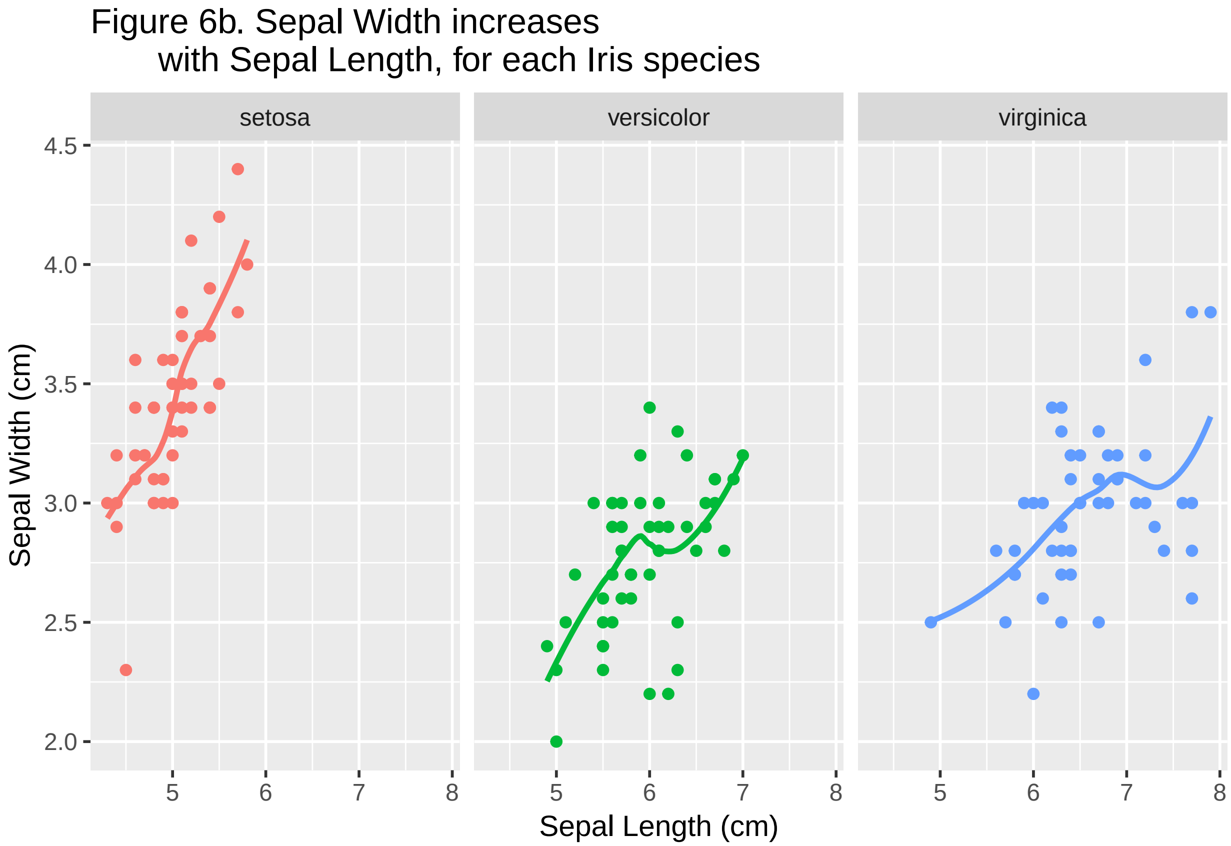
6.1 Plot without Facets (Figure 6a)

```
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)")
```



6.2 Plot with Facets (Figure 6b)

```r
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.

7 Question 7 - Fox or Hedgehog?

We don't write answer sketches for essays.