Stream Macroinvertebrate Analysis Assignment

Bob Hagen

2024-12-05

Table of contents

Introduction-Preparation

First step is to load the "tidyverse" bundle of packages into your R project, which will enable you to run the graphing and summarizing functions.

```
library(tidyverse)
library(readr)
```

If you haven't already done this, you'll also want to enable the "native pipe" operator (|>) in RStudio. To do that, follow these steps:

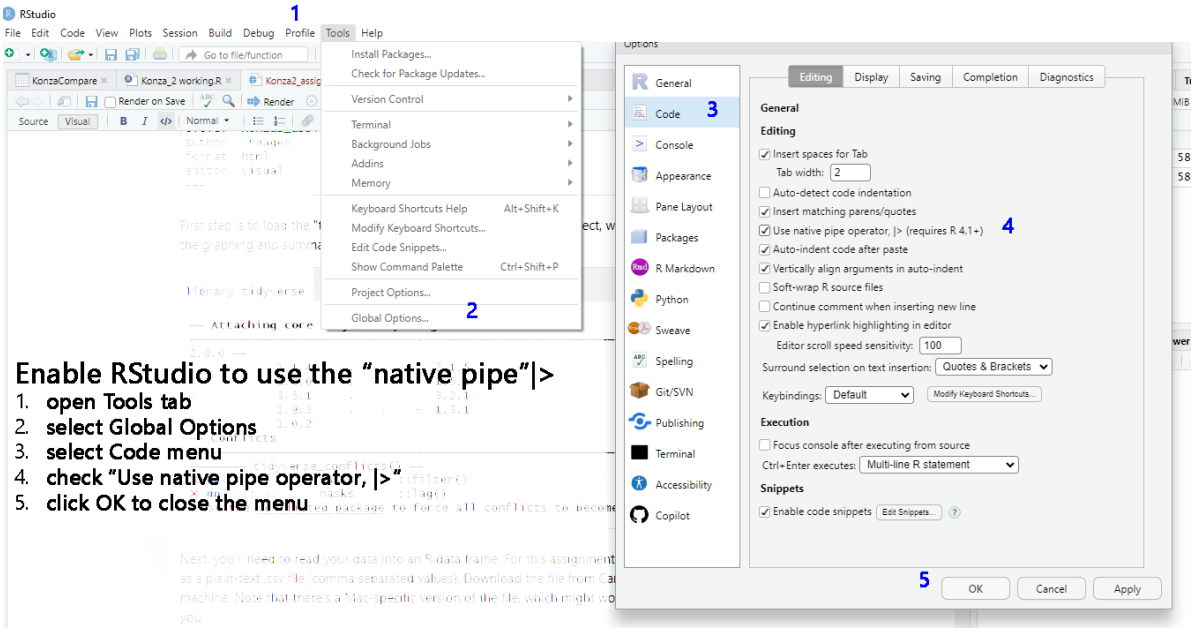


Figure 1: *How to enable RStudio to use the |> operator*

Getting your data into R

Next, you'll need to read your data into an R data frame. For this assignment, you'll need to import 2 data tables after you have finished calculating summary statistics from the 2024 Wakarusa R macroinvertebrate samples **(Part 1 of the assignment)**.

If you have saved the tables as *plain-text .csv* files (comma separated values), you can use the following steps for getting them into R. (Instructions for importing the tables directly from Excel were provided in the first R assignment–note that you will need to install a different package into R, which doesn't work properly if you are using the cloud app!)

If you're using RStudio on your own machine, select the "Import Dataset" tab on the upper right "Environment" window. If you are using the RStudio (Posit) cloud platform, you will first need to upload the files to your project (tab in the lower-right window), before the Import Dataset command will work.

From the dropdown menu, I suggest you use the "From text (readr)" option – that's a more user-friendly alternative. In the popup window, navigate to the data file, select it, then enter a short, descriptive name for the dataset in the lower right "Name" box.

(If you want to use the template codes without having to change the data name, you should name the first one "**Wakarusa**" and the 2nd, "**CompareStreams**".

The dialog box also allows you to specify the format for each column variable. For the Wakarusa table, I chose to save the the ***year*** column as an integer. (As a practical matter, it doesn't matter in this case–just an example here!)

But more usefully, I also wanted the ***area*** variable to be used as a factor, with 2 levels ("upstream" and "downstream"). In the Konza data assignment, we ran a block of R code to convert variables for burn frequency and bison grazing into factor levels. But it's easier to do that when the file is imported by using the read.csv dialog box!!

Repeat the import steps for the 2nd site comparison data table. For this table, change the *site* variable to a factor, with the 4 stream names as levels. When you enter those levels, be sure they exactly match what's in the table—including capital letters.

Screen shots of the process for changing a variable to a factor during the import process

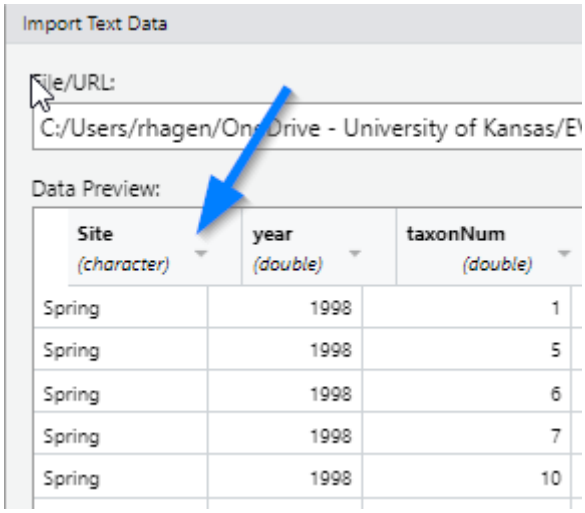


Figure 2: column data type shown here-select the menu

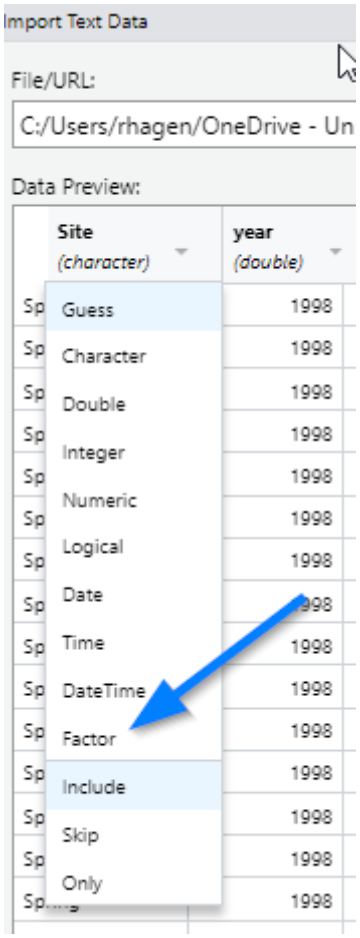


Figure 3: choose “factor” from the dropdown menu

Then enter the levels into the pop up box:

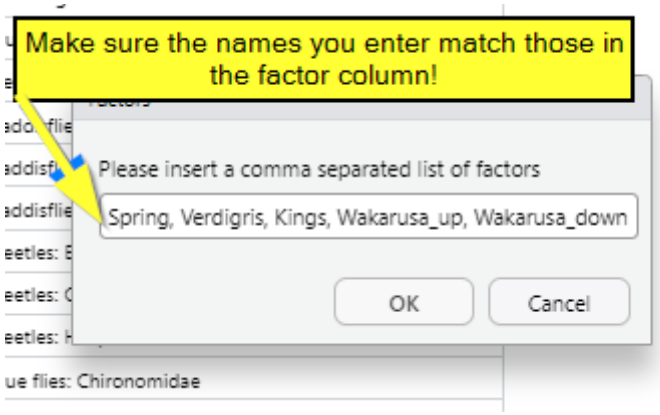


Figure 4: when you've typed them all, choose OK, and finish the importing process!

Examine the data frames to familiarize yourself with the variables and treatment labels. In RStudio, you can inspect the data frames in the Source window, or use the **summary( )** function

Wakarusa Stream Sample Comparison

In this section we'll focus on 2 questions based on the Wakarusa stream invertebrate samples:

- **Do the upstream and downstream sites differ in the 4 measures of diversity and habitat quality?**
- **Do these measures show any consistent change over the 12 years that Field Ecology students have been sampling this site?**

The ***count*** column in the data frame contains the number of invertebrates included in the sample. It's a measure of the collecting effort, so unlike the other summary statistics, *count* isn't really a measure of site conditions. It's possible that differences among samples might just be a result of differences in how much collecting was done each time. (Consider an extreme example: if only 3 invertebrates were collected in total, it's impossible to have richness > 3 taxa for that sample!) It's important to determine how strongly variation in collecting effort affects each of the measures of diversity or habitat quality!

Your first task is to check for relationships between "count"– the number of invertebrates collected – and the other 4 summary statistics.

How much variation in collecting effort was there among the 12 Wakarusa samples? An easy way to find out is with the summary() function. Because that info is only in the *count* variable, you can limit the output to that column alone by using the $ operator: **summary(**<*name of your data frame*>**$count)**

```
summary(Wakarusa$count)
```

1. **What was the range of the *count* variable among the Wakarusa samples (minimum, maximum)?**

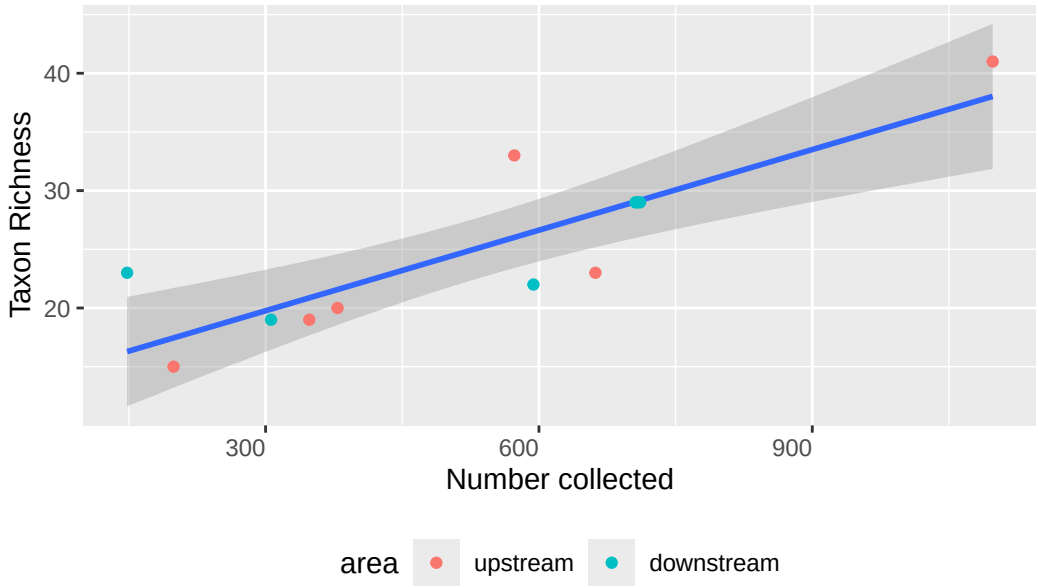
Correlations–scatterplots and trendlines

Scatterplots are a simple way to examine relationships between 2 variables. In the ggplot2 package, that's done with the **geom_point( )** function. Another function, **geom_smooth( )**, helps to reveal general trends within the data as a whole. We can combine those 2 in a single graph with this chunk of code, which plots count on the x-axis and # of taxa (richness) on the y-axis:

```
# in this code chunk, the data are indicated by dots,
# with a shaded band to help
# visualize the relationship

Wakarusa |>
  ggplot(aes(x = count, y = richness,
             )) +
  geom_smooth(method = "lm") +
  geom_point(aes(color = area)) +
  labs(title = "", x = "Number collected", y = "Taxon Richness") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1))
```

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




```
# note that the order of the functions matters:
# by entering geom_point() after the # geom_smooth(),
# the data point dots are drawn on top of the shaded band--
# not hidden below it!
```

The line represents the best estimate of the relationship between the number of taxa and the number of invertebrates collected in this set of samples. The shaded area around the line represents the 95% Confidence Interval around that estimate. (In this case, we used the "method = lm" parameter for geom_smooth, which specified a linear model for the relationship. Other models are possible, but this is a good, simple, approach.)

To interpret the graph, first note the width of the shaded band. A relatively narrow band in this graph implies that the line estimate is a good summary of the relationship, in the same way that a narrow 95% Confidence Interval around a mean implies that it is a good estimate. Second, note the slope of the line. A relatively flat line, or one in which the upper and lower boundaries of the shaded band on the left and right of the graph overlap, indicates no correlation between the two variables.

(To gauge this by eye: if you can draw a horizontal line from the left to right ends of the graph which lies entirely within the gray band, there's no evidence for correlation.)

Here, for the 12 Wakarusa samples, there is a strong–positive–correlation between the collecting effort (number of invertebrates collected) and the taxon richness. Therefore we must be cautious in our interpretation. Differences in taxon richness may not be "real" – instead, they

could be an artifact of different collection effort. (We can adjust for that by correcting for "count" in our analysis.

What is the relationship for the other 3 measures? (Shannon diversity, Percent EPT, Percent Sensitive Taxa). To create these additional graphs, you can copy the same code as above, then change the **y = <variable>** parameter in the **ggplot** and the **labs** lines to the other variables

2. **Create a graph showing the relationship between count and Shannon diversity (ShannonD) for the Wakarusa samples.**
3. **Create a similar graph for percent Ephemeroptera+Plecoptera+Trichoptera (PctEPT) vs. count.:**
4. **Create a graph for percent sensitive taxa (PctSensitive):**
5. **What can you conclude about the relationship in each case?**

Optional question: Are these measures of diversity and stream health independent of each other? If you are interested, try graphing PctEPT vs. ShannonD. In principle, they should capture different aspects of the sample and should not be correlated.

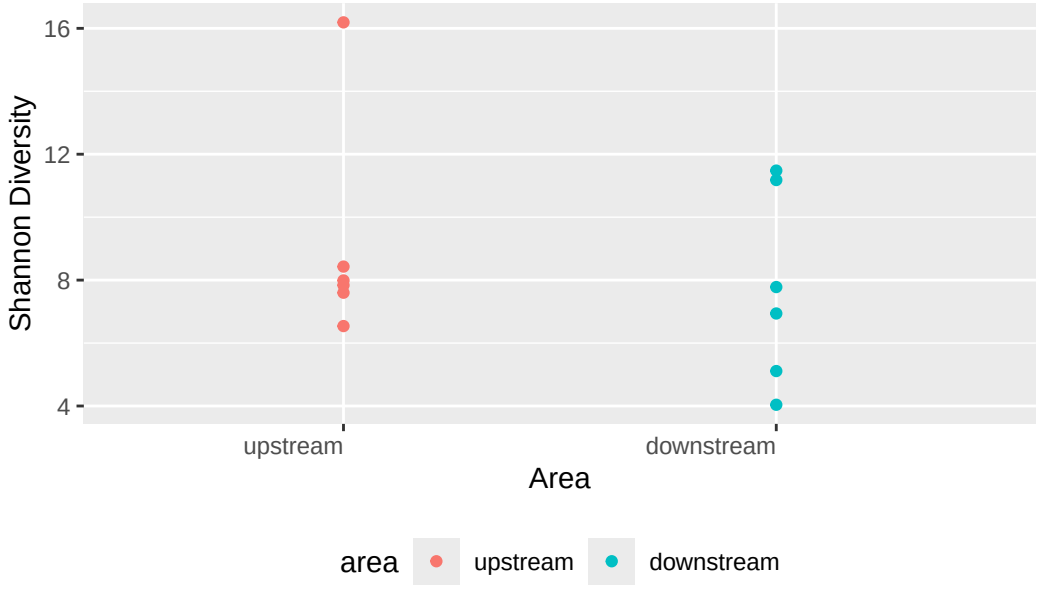
Comparing upstream and downstream Wakarusa samples

Because we have relatively few Wakarusa samples (6 each from upstream and downstream of the waterfall), a boxplot is not very useful. But we can make a simple point graph to display the values directly

Because taxon richness is correlated with collecting effort (count), a comparison graph won't be as useful for that measure. However, informally, we can note that the values for upstream and downstream samples in the first graph we made seem to be intermixed. Samples aren't clustered by area, suggesting that there's no real difference between them with respect to taxon richness.

We can make point graphs for the other 3 measures. Here's code for a graph to compare Shannon Diversity between the areas:

```r
Wakarusa |>
  ggplot(aes(x = area, y = ShannonD,
             color = area)) +
    geom_point() +
  labs(title = "", x = "Area", y = "Shannon Diversity") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1))
```



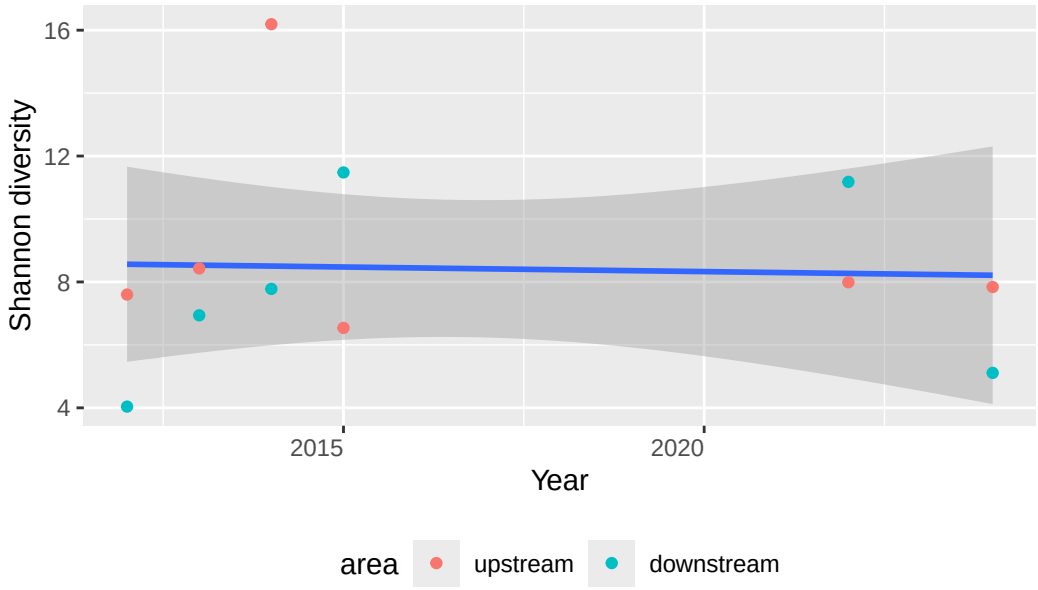
6. **Reuse this code template to create similar graphs for Percent EPT and Percent Sensitive Taxa:**
7. **How do the upstream and downstream samples compare for these 3 measures?**

Comparison among years

Is there evidence of a trend for any of the measures over time? Focus on the 3 measures other than richness.

```
Wakarusa |>
  ggplot(aes(x = year, y = ShannonD,
             )) +
   geom_smooth(method = "lm") +
   geom_point(aes(color = area)) +
  labs(title = "", x = "Year", y = "Shannon diversity") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1))
```

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



8. **Re-use this code to create similar time-series graphs for Percent EPT and Percent Sensitive Taxa**

9. **Is there evidence for linear change over the 12 years spanned by this set of samples?**

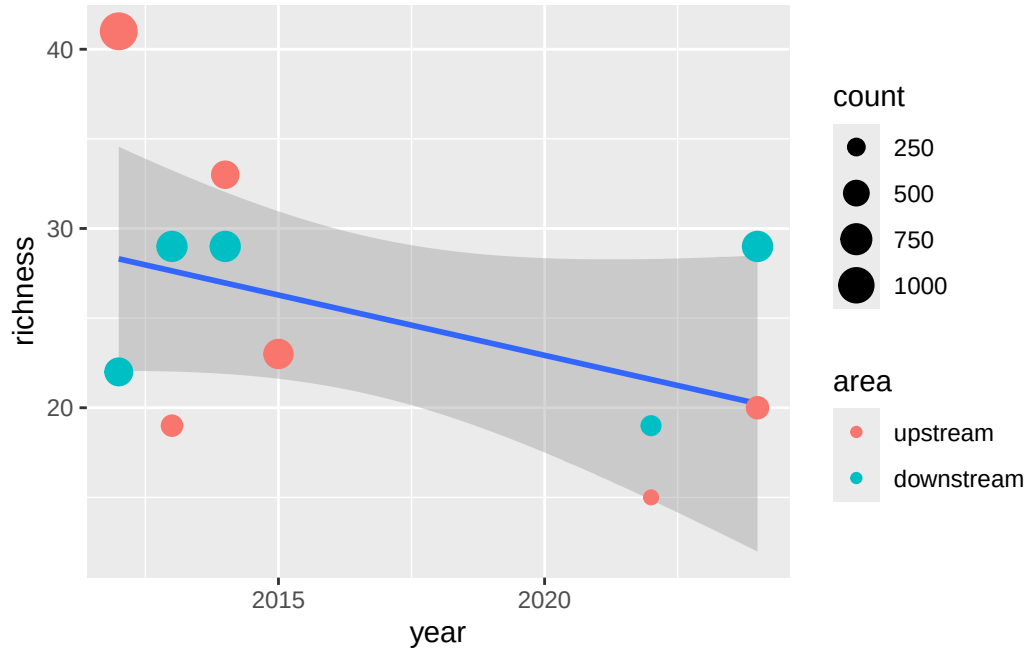
For completeness, we can also graph taxon richness over time. But for this graph, we'll also add an indicator of the collecting effort for each sample. We can do that by adding a new parameter to the **geom_point** function: *size = count* This scales the the size of the points to vary according to the value of *count* for each sample. This can be help reveal potential complications! For example, imagine more invertebrates were typically collected in samples

early in the series. Because of the correlation with taxon richness, those samples would also have had more taxa than later years, regardless of whether the stream's richness had actually changed over time.

Go ahead and run this code (changing the name of the Wakarusa data frame if yours is different):

```
#|output: false
Wakarusa |>
  ggplot(aes(x = year, y = richness,
             )) +
   geom_smooth(method = "lm") +
   geom_point(aes(color = area, size = count))
```

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




```
  labs(title = "", x = "Year", y = "Taxon Richness") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1))
```

```
NULL
```

The graph should (if done correctly!) show a weak negative slope to richness over time. Is this a real change? First, notice that the highest richness occurred in the upstream sample from 2012 (upper left), which is also the largest dot: that sample had the highest value for *count*, indicating it was also the largest collection. In contrast, the lowest richness was seen in the upstream sample from 2022–which was also the smallest sample. It's reasonable to infer that had the 2022 sample included more invertebrates, it would have had more taxa. Because we have relatively few samples, these two extreme points alone can tilt the curve downward to the right! It's likely that the slope of the estimate line is just an artifact.

That interpretation is reinforced by the wide confidence interval. A completely horizontal estimate line (or even one with a slight upward slope)! would fit within the shaded band. Be cautious in drawing conclusions from relatively small data sets.

10. **What overall conclusions can you draw from the graphs for these measures among the Wakarusa samples? Consider the potential effects of both year and area (upstream vs downstream).**

Comparison of Wakarusa Samples with Reference Stream sites

Now we'll examine the larger set of samples Our goal is to compare the Wakarusa samples collected by students with samples from some eastern Kansas reference stream sites collected by agency employees. Reference sites are those that experienced stream biologists have identified as representing the best conditions for that region.

- **How does the Wakarusa compare with the 3 reference stream sites with respect to these measures of stream health?**
- **Did samples collected from 2 reference sites (Spring River and Verdigris River) show a pattern of variation among years (or lack of variation) similar to the Wakarusa samples?**

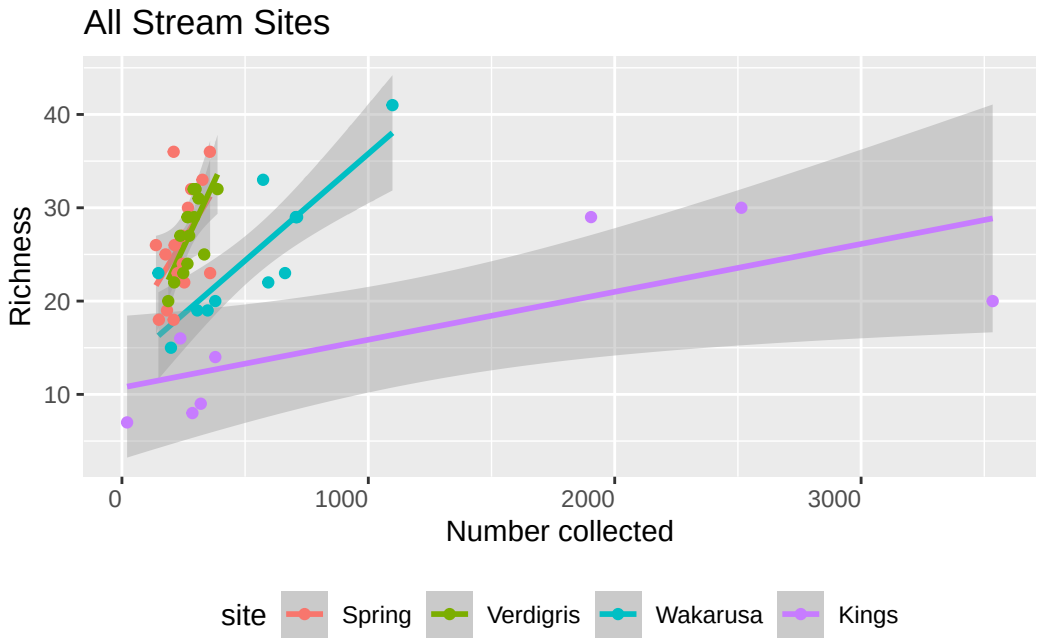
  In other words: **How consistent are these measures of invertebrate diversity and stream health over time in the absence of obvious disturbance?**

Correlations with *count* (collecting effort): all sites

First, for this larger set, we should look at the relationship of count– our measure of collecting effort–with the 4 measures of stream health as we did for the Wakarusa samples. Here is a code for a graph of the relationship between count and richness for the *CompareStreams* data frame with all samples.

```
CompareStreams |>
  ggplot(aes(x = count, y = richness,
             color = site)) +
  geom_smooth(method = "lm") +
  geom_point( ) +
  labs(title = "All Stream Sites", x = "Number collected", y = "Richness") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1))
```

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



The code is *almost* identical to what we used for the first graph. But by moving the **color =** <factor> parameter from the **geom_point** function up to the **ggplot** function, it applies to both the geom_point and the geom_smooth functions. This means that ggplot2 will calculate a separate estimate line for each of the sites, as well as making the points for each site a different color. (We also added a graph title because we're starting to generate a lot of these figures!)

Run the code and examine the graph. All of the sites show a positive relationship, but differ in both the slope and how tight the correlation is. We should probably set aside taxon richness to avoid this complicating factor!

Modify the code to create graphs for the relationship between count and the other 3 measures.

11. **graph Shannon Diversity vs count:**
12. **graph Percent EPT vs. count:**
13. **Graph Percent sensitive taxa vs count:**

It's a bit hard to see what's going on with the Spring R and Verdigris R samples because they're squeezed into the left side of the graph. The simplest way to improve the display is to plot those sites separately from the sites with much larger counts.

We can create a new subset of the data frame with just those sites and rerun the code on this subset:

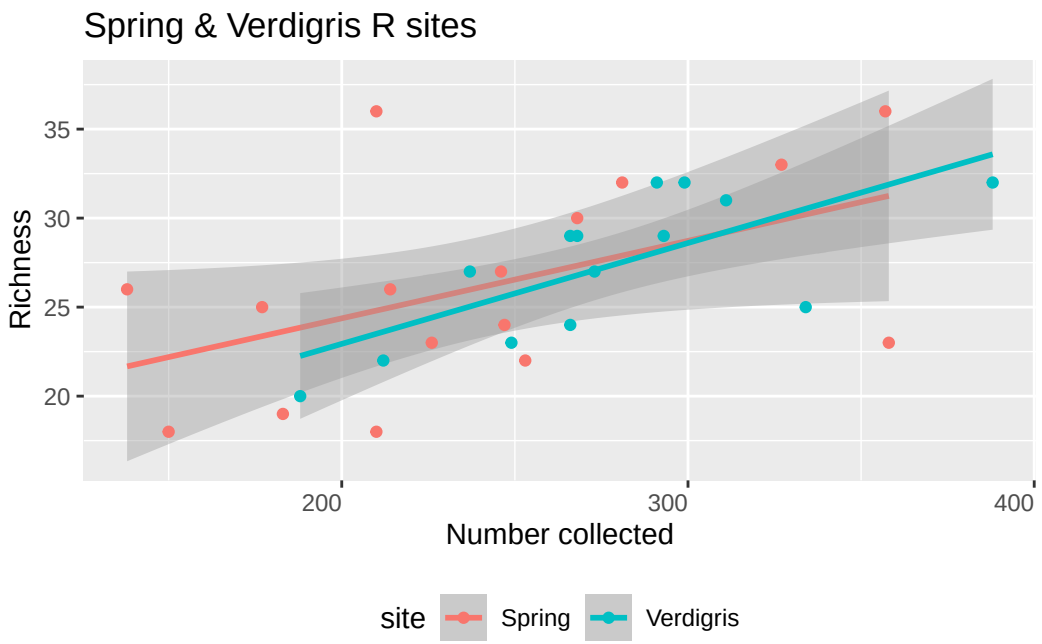
```
SpringVerdigris <- filter(CompareStreams, site == "Spring" | site == "Verdigris" )

 # the == operator means "equals", and the | operator means "or"

view(SpringVerdigris)
```

```
SpringVerdigris |>
  ggplot(aes(x = count, y = richness,
             color = site)) +
  geom_smooth(method = "lm") +
  geom_point( ) +
  labs(title = "Spring & Verdigris R sites", x = "Number collected",
       y = "Richness") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1))
```

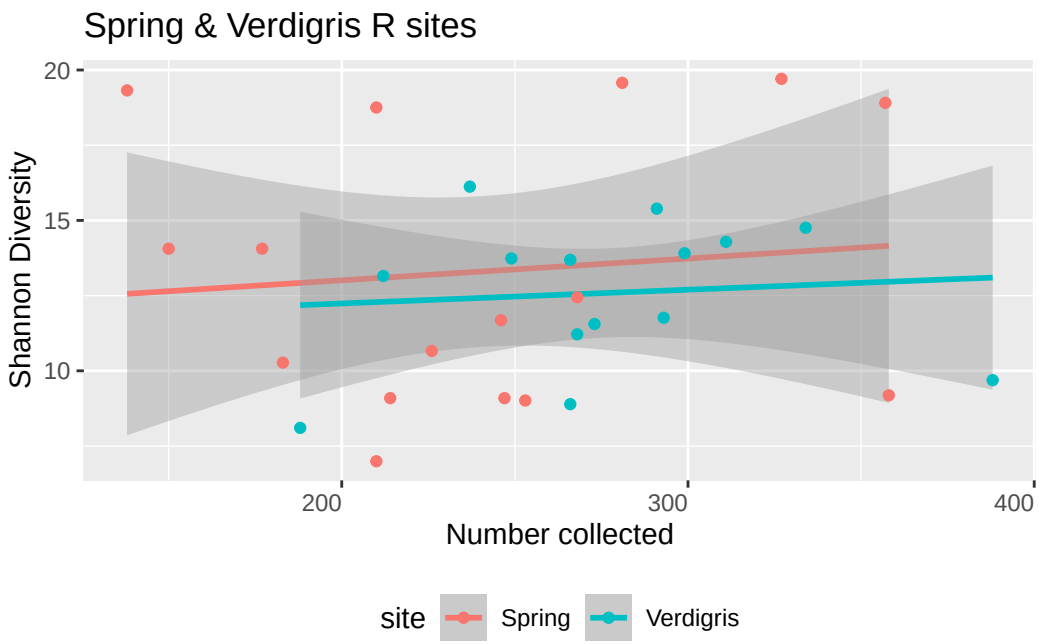
```
`geom_smooth()` using formula = 'y ~ x'
```



We can do the same for ShannonD, PctEPT, and PctSensitive:

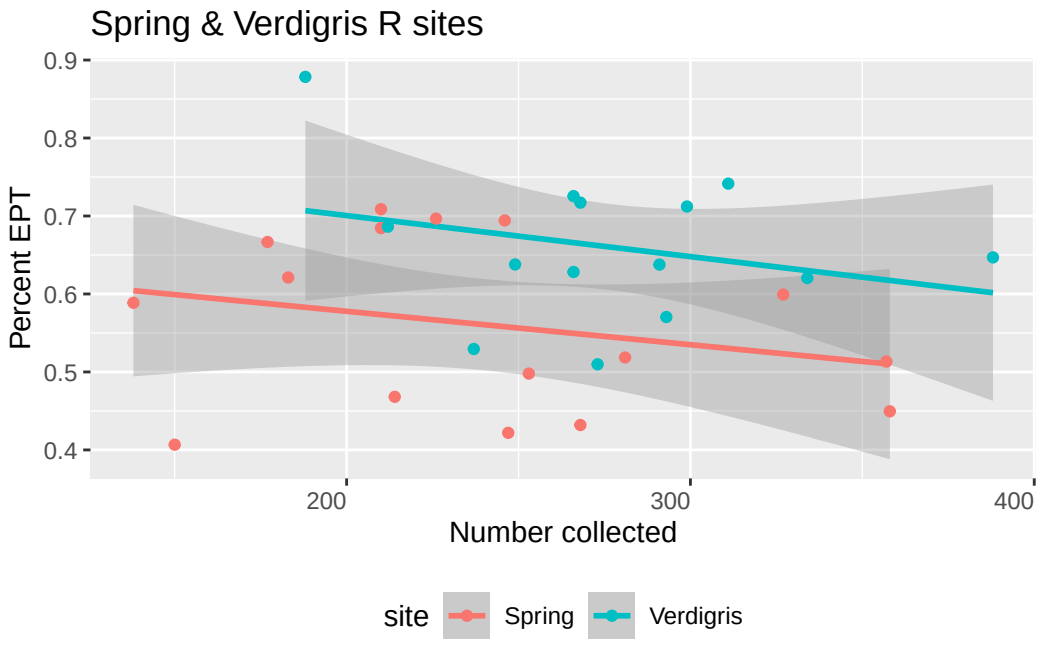
```
SpringVerdigris |>
  ggplot(aes(x = count, y = ShannonD,
             color = site )) +
  geom_smooth(method = "lm") +
  geom_point() +
  labs(title = "Spring & Verdigris R sites", x = "Number collected",
       y = "Shannon Diversity") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1))
```

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



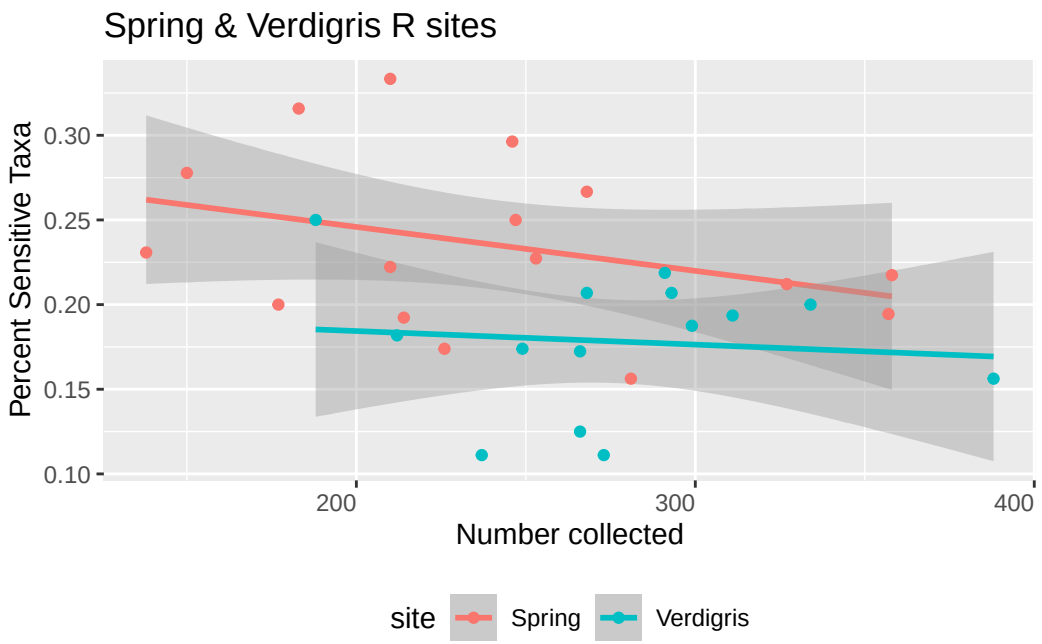

```
SpringVerdigris |>
  ggplot(aes(x = count, y = PctEPT,
             color = site )) +
  geom_smooth(method = "lm") +
  geom_point() +
  labs(title = "Spring & Verdigris R sites", x = "Number collected",
       y = "Percent EPT") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1))
```

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




```
SpringVerdigris |>
  ggplot(aes(x = count, y = PctSensitive,
             color = site )) +
  geom_smooth(method = "lm") +
  geom_point() +
  labs(title = "Spring & Verdigris R sites", x = "Number collected",
       y = "Percent Sensitive Taxa") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1))
```

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



14. **Based on these graphs, which of the measures appear to be uncomplicated by correlation with the collection effort (*count* variable)?**

Optional question: Are these measures of diversity and stream health independent of each other? If you are interested, try graphing PctEPT vs. ShannonD. In principle, they should capture different aspects of the sample and should not be correlated. Run the same graph code as above, this time for the Spring and Verdigris stream sites:

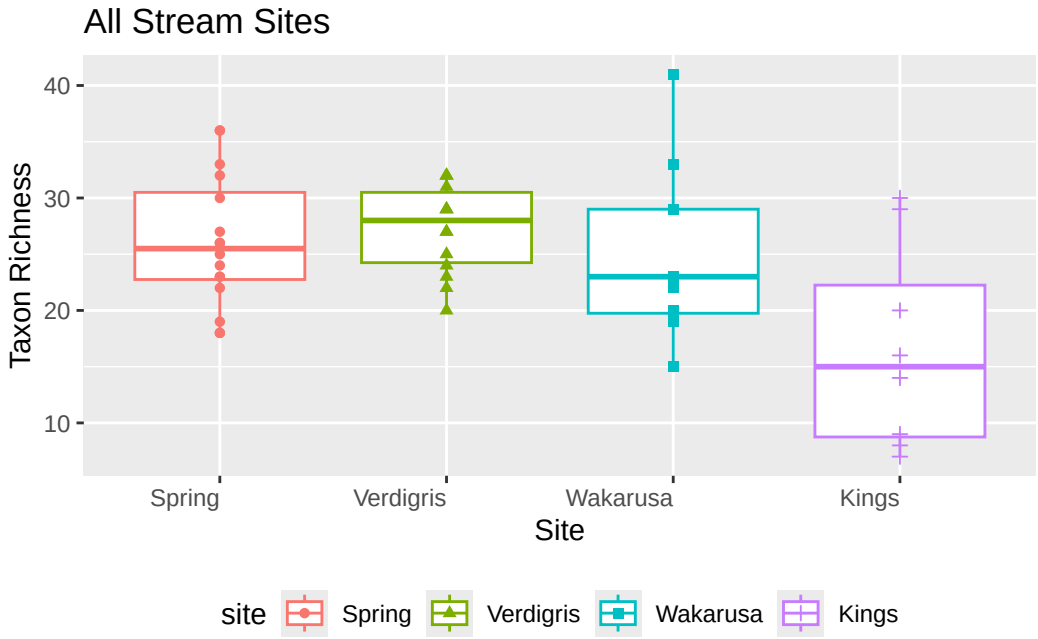
```
SpringVerdigris |>
  ggplot(aes(x = ShannonD, y = PctEPT,
             color = site )) +
  geom_smooth(method = "lm") +
  geom_point() +
  labs(title = "Spring & Verdigris: Correlation", x = "Shannon Diversity",
       y = "Percent EPT") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1))
```

Comparison among sites

A simple way to visualize variation among the 4 stream sites is to use a box plot. Because we have relatively few samples, it's a good idea to include the actual data points too. In this example, each site's points will have both a unique color, which is the same as the boxplot outline because the "*color = site*" parameter is under the **ggplot** function, and a unique shape because "*shape = site*" is under **geom_point**.

boxplot – richness

```
CompareStreams |>
  ggplot(aes(x = site, y = richness,
             color = site )) +
   geom_boxplot() +
   geom_point(aes(shape = site)) +
  labs(title = "All Stream Sites", x = "Site", y = "Taxon Richness") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1))
```



As noted above, interpretation of Taxon Richness is problematic because it's correlated with count. But you can use this code to create graphs for the other 3 measures:.

15. **Comparison among sites: Shannon diversity**

16. **Comparison among sites: Percent EPT**

17. **Comparison among sites: Percentage of sensitive taxa**

18. **Write a brief summary of your interpretation of these 3 boxplot graphs. Specifically, how does the Wakarusa compare to these 3 eastern Kansas reference sites?**

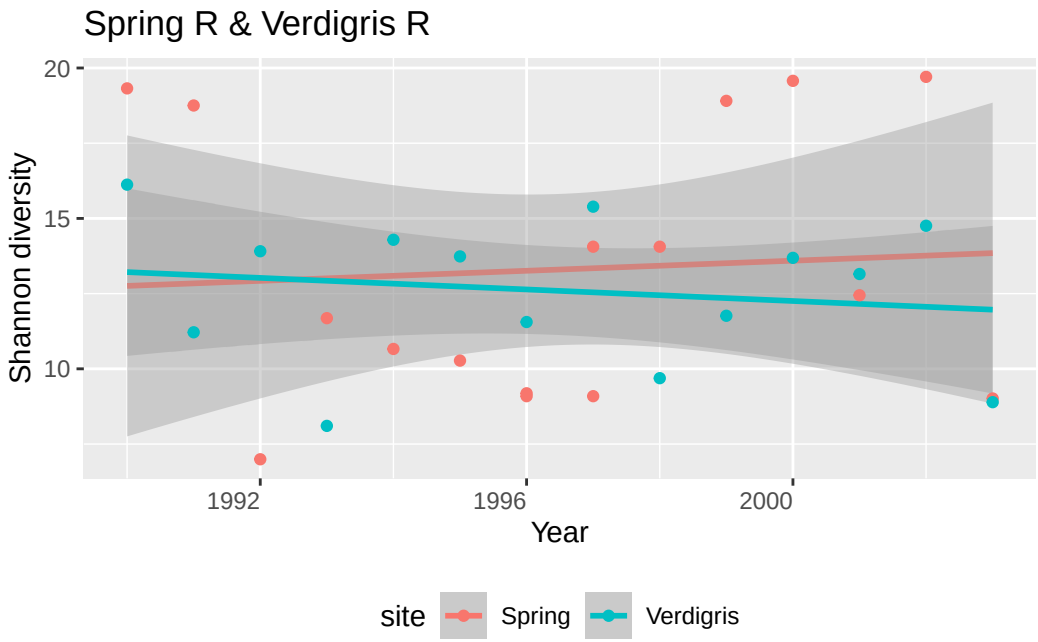
Comparison among years: Spring R and Verdigris R sites

The Kansas Reference stream data set includes 16 samples from the Spring River and 14 samples from the Verdigris River sites (1990 -2003). How does the pattern of variation over years (or lack of variation!) compare with the Wakarusa samples? We'll set aside Taxon Richness, and consider only the other 3 measures.

This block of code will plot Shannon Diversity vs. year for the Spring and Verdigris river sites:

```
SpringVerdigris |>
  ggplot(aes(x = year, y = ShannonD,
             color = site)) +
   geom_smooth(method = "lm") +
   geom_point( ) +
  labs(title = "Spring R & Verdigris R", x = "Year",
       y = "Shannon diversity") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1))
```

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



19. **Spring River and Verdigris River: Percent EPT vs. year.** Modify the code above to create this graph.

20. **Spring River and Verdigris River: Percent Sensitive Taxa vs. year.** Modify the code to create this graph.

Answering the Questions

You've created a number of graphs in the steps above. Your final task is to summarize and assemble these results to answer the larger questions:

With respect to the Wakarusa River stream invertebrate samples they are:

- **Do the upstream and downstream sites differ in diversity or habitat quality?**
- **Do these measures show any consistent change over the 12 years that Field Ecology students have been sampling this site?**

  Then:

- **How does the Wakarusa River compare with the group of 3 reference stream sites, with respect to these measures of stream health?**

- **Did samples collected from 2 reference sites (Spring River and Verdigris River) show a pattern of variation among years (or lack of variation) similar to the Wakarusa samples?**

  In other words: **How consistent are these measures of invertebrate diversity and stream health over time in the absence of obvious disturbance?**

Emphasize the summary statistics (measures) that seem most informative. Create a Word or PDF document with your answers to these 4 questions. Your document must include relevant graphs you've created in R, as well as text and captions that explain those graphs.