BISC 307, Lab 2B: Photosynthesis & Leaf Traits

Prof. Jackie Hatala Matthes, Fall 2018

Lab: 25 September

Lab 2b Objectives

1. Learn to explore patterns in data that can be used to generate hypotheses and examine how ecological and evolutionary processes control photosynthesis.
2. Examine variation within variables, and covariation among variables within a global dataset for leaf carbon exchange.
3. Use models to extract patterns out of data.

1. Introduction

In this lab, we'll work with a global dataset of leaf carbon exchange measurements and leaf traits collected across 626 individual plants of 98 species at 12 sites in North and South America that span 58 degrees of latitude and published in in Smith & Dukes (2017). Our lab objectives for today will practice and build on the sections on Exploratory Data Analysis that you read in preparation for this week's lab, Sections 7.1-7.6 from the R for Data Science book, in addition to exploring global patterns in leaf carbon/water exchange.

Your first step should be to create an R script file and to set your working directory (or to make a Project file). Next, we'll load the libraries that we'll be using.

```r
# Load libraries
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Warning: package 'tibble' was built under R version 3.4.3

## Warning: package 'tidyr' was built under R version 3.4.2

## Warning: package 'purrr' was built under R version 3.4.2

## Warning: package 'dplyr' was built under R version 3.4.2

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats
## lag():    dplyr, stats
```

```r
# Load plantecophys package to fit A-Ci curves
library(plantecophys)
```

2. A-Ci Curve Fitting

The key data for this lab are comprised of the biologically relevant parameters fit with A-Ci curves collected on individuals. The raw data are all within the `LCE_ACi_curves` directory (inside the `data` directory) in the lab folder. We'll work through one example of an A-Ci curve fit, to see how the parameters of interest, Vcmax, Jmax, and Rd, are estimated.

Most of the computational heavy lifting for A-Ci curve fitting is done with nonlinear estimators from the `plantecophys` package, which save us from having to code the three equations ourselves. It is, however, important to conceptually understand what is going on within the A-Ci curve fitting. Remember that in class, we discussed the three equations that are simultaneously being fit for Rubisco-limited photosynthesis (Ac), RuBP-limited photosynthesis (Aj), and stomatal conductance (gs).

```
# Read in example A-Ci curve data
ID <- "Auburn_Iopa_1"
aci_ID <- read_csv(paste0("data/LCE/LCE_ACi_curves/",ID,".csv"))
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   X1 = col_integer(),
##   LA = col_integer(),
##   Pari = col_integer(),
##   id = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
# Fit A-Ci curve with fitaci function from plantecophys package
fit_ID <- fitaci(aci_ID,
                 varnames = list(ALEAF = "Photo",
                                 Tleaf = "Tleaf", Ci = "Ci", PPFD = "Pari"),
                 Tcorrect=TRUE, useRd=FALSE)

# Look at the estimated parameters from the A-Ci curve fitting
fit_ID
```

```
## Result of fitaci.
##
## Data and predictions:
##        Ci  Ameas     Amodel         Ac         Aj   Ap      Rd VPD Tleaf
## 5    52.8 -0.301 -0.8873391  0.2994207  0.7400495 1000 1.18674 1.5 26.14
## 4    86.1  0.611  0.3708781  1.5577624  3.2336653 1000 1.18674 1.5 26.41
## 3   144.0  1.950  2.3808266  3.5680987  5.9576303 1000 1.18674 1.5 26.36
## 2   197.0  3.010  4.0581210  5.2460740  7.5119160 1000 1.18674 1.5 26.69
## 1   249.0  4.160  5.5670094  6.7562259  8.5952781 1000 1.18674 1.5 26.87
## 7   287.0  7.490  6.4199929  7.6105870  9.1078649 1000 1.18674 1.5 26.36
## 6   301.0  7.200  6.7150696  7.9063859  9.2660556 1000 1.18674 1.5 26.22
## 8   416.0  9.850  9.1092856 10.3959054 10.4021604 3000 1.18674 1.5 26.27
## 9   495.0 10.700  9.7800662 11.9575859 10.9789448 3000 1.18674 1.5 26.46
## 10  546.0 10.300 10.0704232 12.8485172 11.2651260 3000 1.18674 1.5 26.51
## 11  577.0  9.740 10.2465164 13.4073366 11.4398777 3000 1.18674 1.5 26.60
## 12  673.0  9.940 10.6172628 14.7942194 11.8086620 3000 1.18674 1.5 26.58
## 13  917.0 11.000 11.2018872 17.5149207 12.3916207 3000 1.18674 1.5 26.46
##            Cc PPFD Patm Ci_original
## 5    52.79911 1801  100        52.8
## 4    86.10037 1799  100        86.1
## 3   144.00238 1799  100       144.0
## 2   197.00406 1800  100       197.0
## 1  249.00557 1801  100       249.0
## 7  287.00643 1801  100       287.0
## 6  301.00672 1801  100       301.0
## 8  416.00912 1799  100       416.0
## 9  495.00979 1800  100       495.0
## 10 546.01008 1799  100       546.0
## 11 577.01026 1799  100       577.0
## 12 673.01063 1799  100       673.0
## 13 917.01121 1799  100       917.0
## 
## Root mean squared error:  2.692399 
## 
## Estimated parameters:
##       Estimate Std. Error
## Vcmax 29.45043  3.2374570
## Jmax  55.07919  3.4871049
## Rd     1.18674  0.6226442
## Note: Vcmax, Jmax are at 25C, Rd is at measurement T.
## 
## Curve was fit using method:  default 
## 
## Parameter settings:
## Patm = 100 
##  alpha = 0.24 
##  theta = 0.85 
##  EaV = 82620.87 
##  EdVC = 0 
##  delsC = 645.1013 
##  EaJ = 39676.89 
##  EdVJ = 2e+05 
##  delsJ = 641.3615 
## 
## Estimated from Tleaf (shown at mean Tleaf):
## GammaStar =  46.04335 
## Km =  805.5637
```

Within the output from the `fitaci()` function, we can see the variables that we are most interested in, Vcmax, Jmax, and Rd. For the rest of lab, we are going to work with a dataset where these parameters have already been calculated from the 626 A-Ci curves aggregated within this global dataset.

3. Global dataset of Vcmax, Jmax, and Rd

This global dataset used A-Ci curved collected from plants within ecosystems in North and South America to find Vcmax, Jmax, and Rd. The rest of our lab will focus on examining patterns, generating questions, and testing hypotheses for the values of Vcmax, Jmax, and Rd that have already been calculated in the `LCE_data.csv` file (LCE = leaf carbon exchange). First, let's load the data and see what it looks like:

```
LCE_global <- read_csv("data/LCE/LCE_data.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   index = col_integer(),
##   Date = col_character(),
##   Year = col_integer(),
##   Day = col_integer(),
##   Location = col_character(),
##   Genus = col_character(),
##   Species = col_character(),
##   Rep = col_integer(),
##   Phenology = col_character(),
##   Lifespan = col_character(),
##   Photosynthesis = col_character(),
##   Stature = col_character(),
##   Juvenile = col_character(),
##   FCper = col_integer(),
##   Satper = col_integer(),
##   aci_id = col_character(),
##   leaf_shape = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
LCE_global
```

```
## # A tibble: 598 x 131
##    index Date   Year   Day Location   Lat   Lon Genus Species   Rep
##    <int> <chr> <int> <int> <chr>    <dbl> <dbl> <chr> <chr>   <int>
##  1     1 5/19~  2015   139 LaSelva   10.4 -84.0 Albi~ adinoc~     1
##  2     2 5/21~  2015   141 LaSelva   10.4 -84.0 Alch~ costar~     1
##  3     3 5/22~  2015   142 LaSelva   10.4 -84.0 Anno~ papili~     1
##  4     4 5/23~  2015   143 LaSelva   10.4 -84.0 Anno~ papili~     2
##  5     5 5/23~  2015   143 LaSelva   10.4 -84.0 Anno~ papili~     3
##  6     6 5/26~  2015   146 LaSelva   10.4 -84.0 Anno~ papili~     4
##  7     7 5/26~  2015   146 LaSelva   10.4 -84.0 Anno~ papili~     5
##  8     8 5/26~  2015   146 LaSelva   10.4 -84.0 Anno~ papili~     6
##  9     9 5/27~  2015   147 LaSelva   10.4 -84.0 Anno~ papili~     7
## 10    10 5/24~  2015   144 LaSelva   10.4 -84.0 Apei~ membra~     1
## # ... with 588 more rows, and 121 more variables: Phenology <chr>,
## #   Lifespan <chr>, Photosynthesis <chr>, Stature <chr>, Juvenile <chr>,
## #   Vcmax <dbl>, Jmax <dbl>, Vpmax <dbl>, Rd <dbl>, Tleaf_photo <dbl>,
## #   Tleaf_R <dbl>, LA <dbl>, LM <dbl>, CN <dbl>, Nper <dbl>, Cper <dbl>,
## #   LWP_MPa <dbl>, DBH <dbl>, Height <dbl>, SMper <dbl>, FCper <int>,
## #   Satper <int>, Tavg1 <dbl>, Tavg2 <dbl>, Tavg3 <dbl>, Tavg4 <dbl>,
## #   Tavg5 <dbl>, Tavg6 <dbl>, Tavg7 <dbl>, Tavg8 <dbl>, Tavg9 <dbl>,
## #   Tavg10 <dbl>, Tavg11 <dbl>, Tavg12 <dbl>, Tavg13 <dbl>, Tavg14 <dbl>,
## #   Tavg15 <dbl>, Tavg16 <dbl>, Tavg17 <dbl>, Tavg18 <dbl>, Tavg19 <dbl>,
## #   Tavg20 <dbl>, Tavg21 <dbl>, Tavg22 <dbl>, Tavg23 <dbl>, Tavg24 <dbl>,
## #   Tavg25 <dbl>, Tavg26 <dbl>, Tavg27 <dbl>, Tavg28 <dbl>, Tavg29 <dbl>,
## #   Tavg30 <dbl>, Tavg31 <dbl>, Tavg32 <dbl>, Tavg33 <dbl>, Tavg34 <dbl>,
## #   Tavg35 <dbl>, Tavg36 <dbl>, Tavg37 <dbl>, Tavg38 <dbl>, Tavg39 <dbl>,
## #   Tavg40 <dbl>, Tavg41 <dbl>, Tavg42 <dbl>, Tavg43 <dbl>, Tavg44 <dbl>,
## #   Tavg45 <dbl>, Tavg46 <dbl>, Tavg47 <dbl>, Tavg48 <dbl>, Tavg49 <dbl>,
## #   Tavg50 <dbl>, Tavg51 <dbl>, Tavg52 <dbl>, Tavg53 <dbl>, Tavg54 <dbl>,
## #   Tavg55 <dbl>, Tavg56 <dbl>, Tavg57 <dbl>, Tavg58 <dbl>, Tavg59 <dbl>,
## #   Tavg60 <dbl>, Tavg61 <dbl>, Tavg62 <dbl>, Tavg63 <dbl>, Tavg64 <dbl>,
## #   Tavg65 <dbl>, Tavg66 <dbl>, Tavg67 <dbl>, Tavg68 <dbl>, Tavg69 <dbl>,
## #   Tavg70 <dbl>, Tavg71 <dbl>, Tavg72 <dbl>, Tavg73 <dbl>, Tavg74 <dbl>,
## #   Tavg75 <dbl>, Tavg76 <dbl>, Tavg77 <dbl>, Tavg78 <dbl>, ...
```

One of the first things to do when you start exploring a big dataset like this that you haven't seen before is to understand the dimensions of the data (which we can see from looking at the tibble above), and what type of replication there is within the major cateogires of the dataset. This can help to inform what to do next with the data.

```
# Find numbers of "replicates" at each Locations (site)
LCE_global %>%
  group_by(Location) %>%
  summarize(count = n())
```

```
## # A tibble: 12 x 2
##    Location   count
##    <chr>      <int>
##  1 Auburn        57
##  2 Blandy        58
##  3 BNZ           54
##  4 Fermi         21
##  5 GR            45
##  6 Guelph        47
##  7 Horizontes    66
##  8 KBS           40
##  9 LaSelva       87
## 10 MMSF          52
## 11 SEPAC         27
## 12 UMBS          44
```

```
# Find numbers of "replicates" for C3 and C4 plants
LCE_global %>%
  group_by(Photosynthesis) %>%
  summarize(count = n())
```

```
## # A tibble: 2 x 2
##   Photosynthesis count
##   <chr>          <int>
## 1 C3               555
## 2 C4                43
```

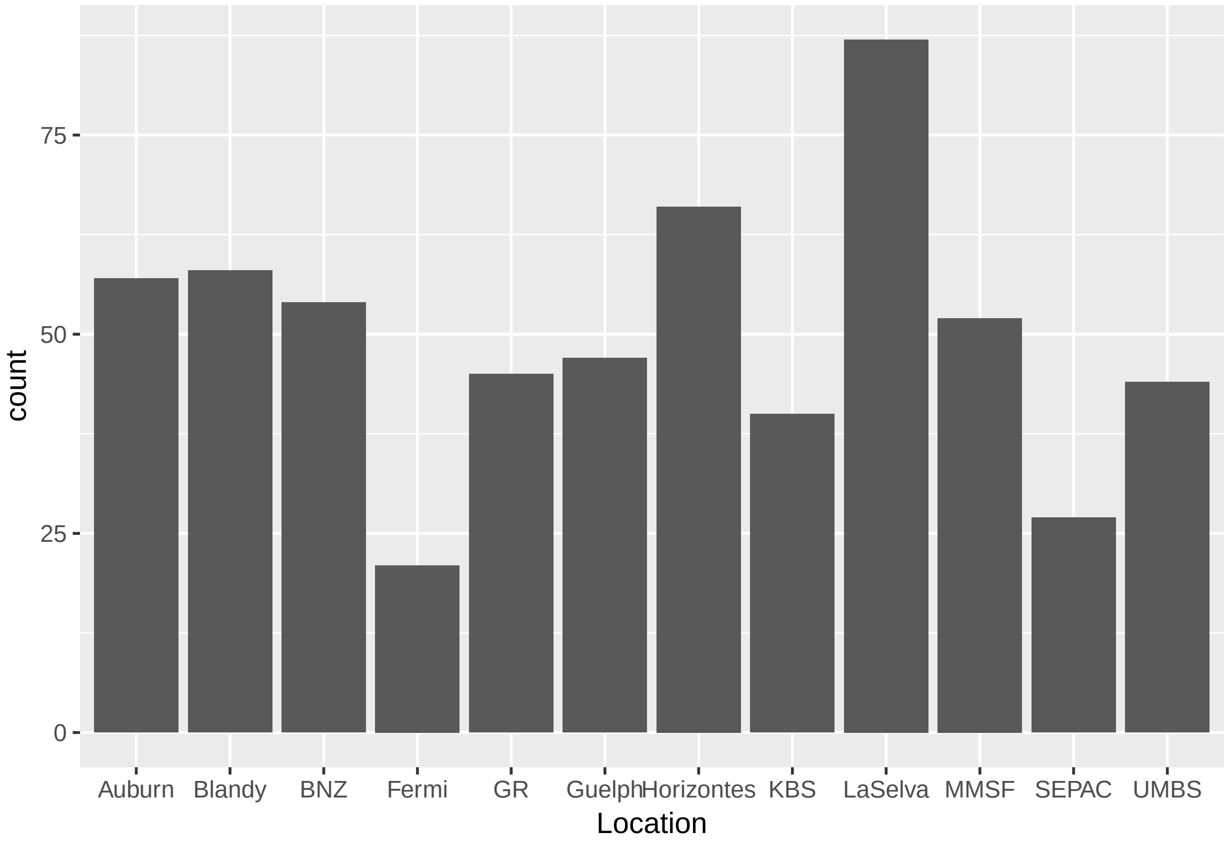
Code Challenge 1: Examine numbers of replicates within at least two additional variables that you might be interested in examining further. Remember that the potential variables for exploration are all the column names of the LCE_global tibble.

---

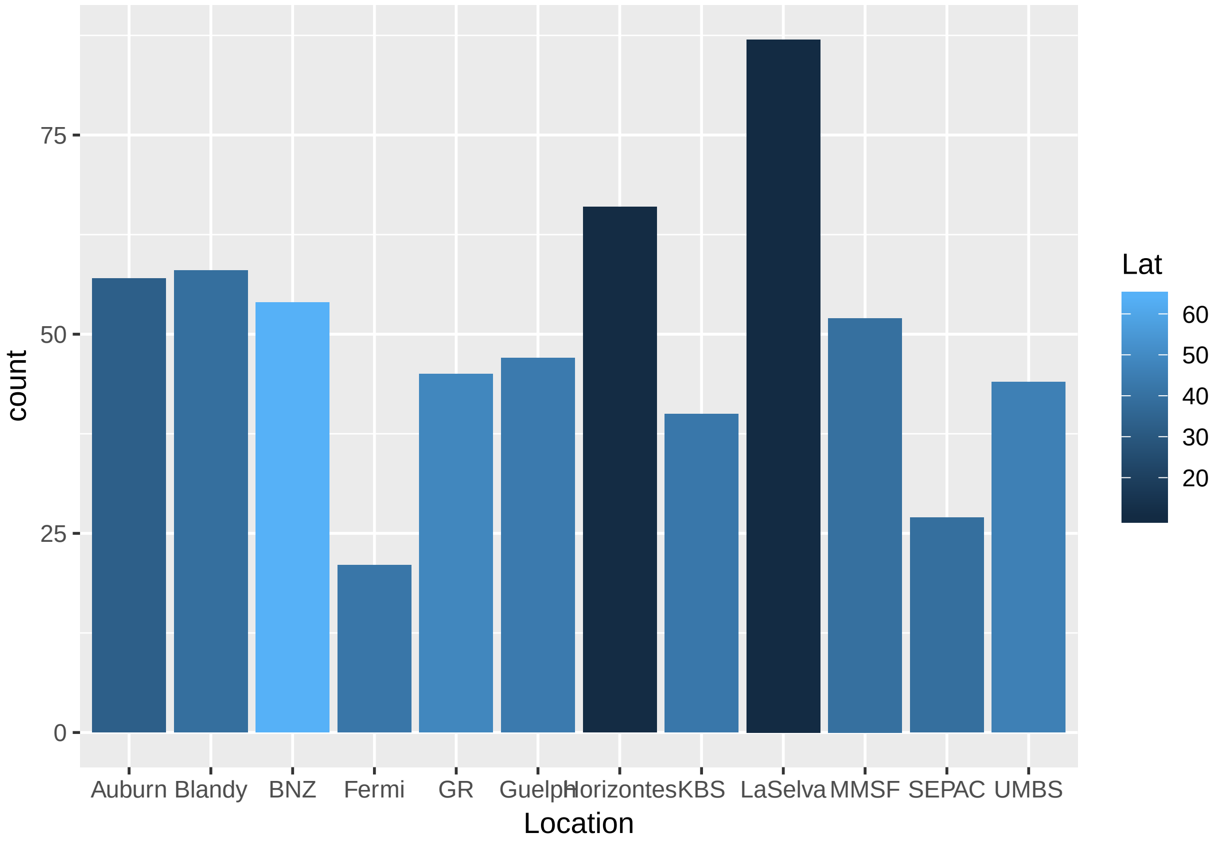
4. Visualizing variation within a single variable & covariation among a categorical and continuous variable

The first place to start when you are doing exploratory data analysis is to visualize the distribution of your data. For categorical data, this is what we did when we generated "count" tables above. The tables are telling us how our dataset is distributed among those categories. We also could have plotted this using geom_bar:

```
# Plot data by Location
ggplot(data = LCE_global) +
  geom_bar(mapping = aes(x = Location))
```



```r
# Plot data by Location - add color to see which latitude sites are at
ggplot(data = LCE_global) +
  geom_bar(mapping = aes(x = Location, fill = Lat))
```

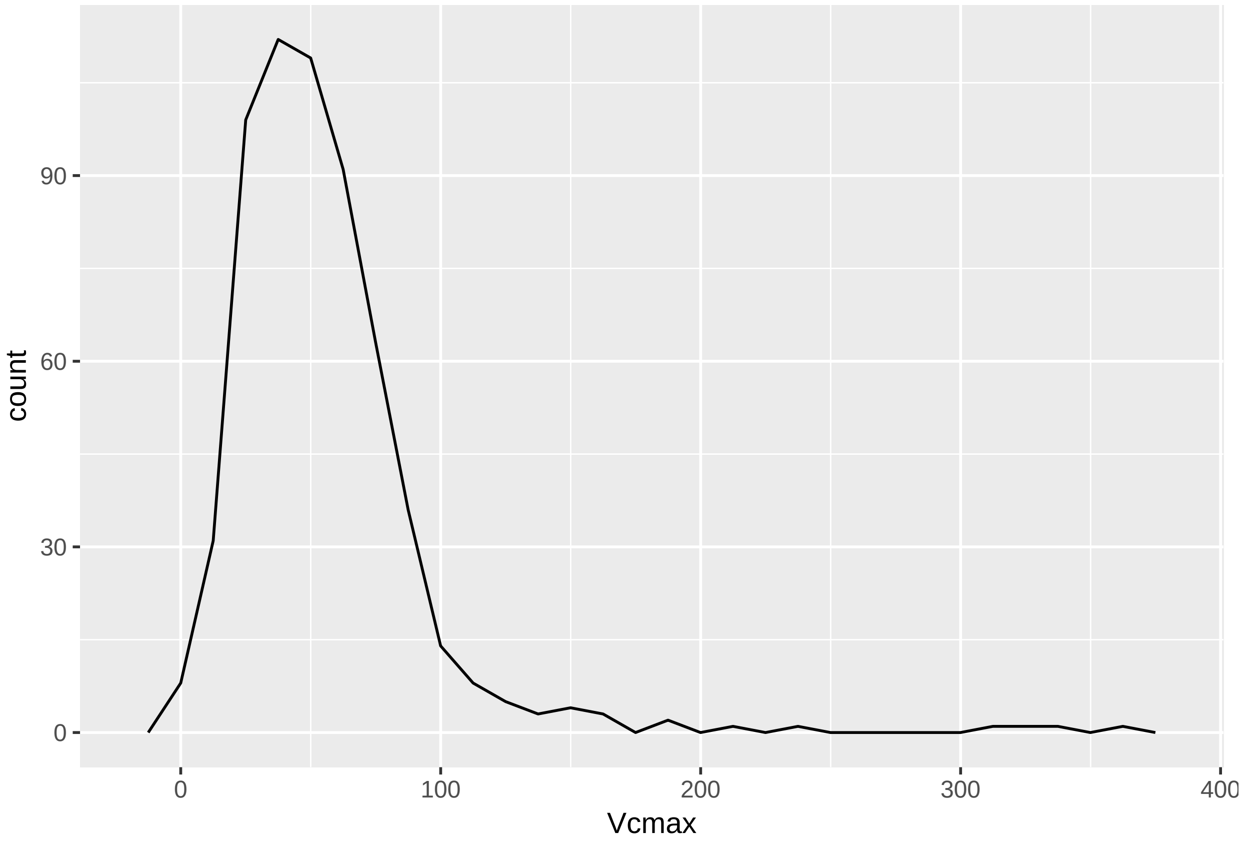


We can also examine the distribution of our continuous variables with geom_freqpoly:

```
# Plot Vcmax distribution
ggplot(data = LCE_global) +
  geom_freqpoly(aes(x = Vcmax))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

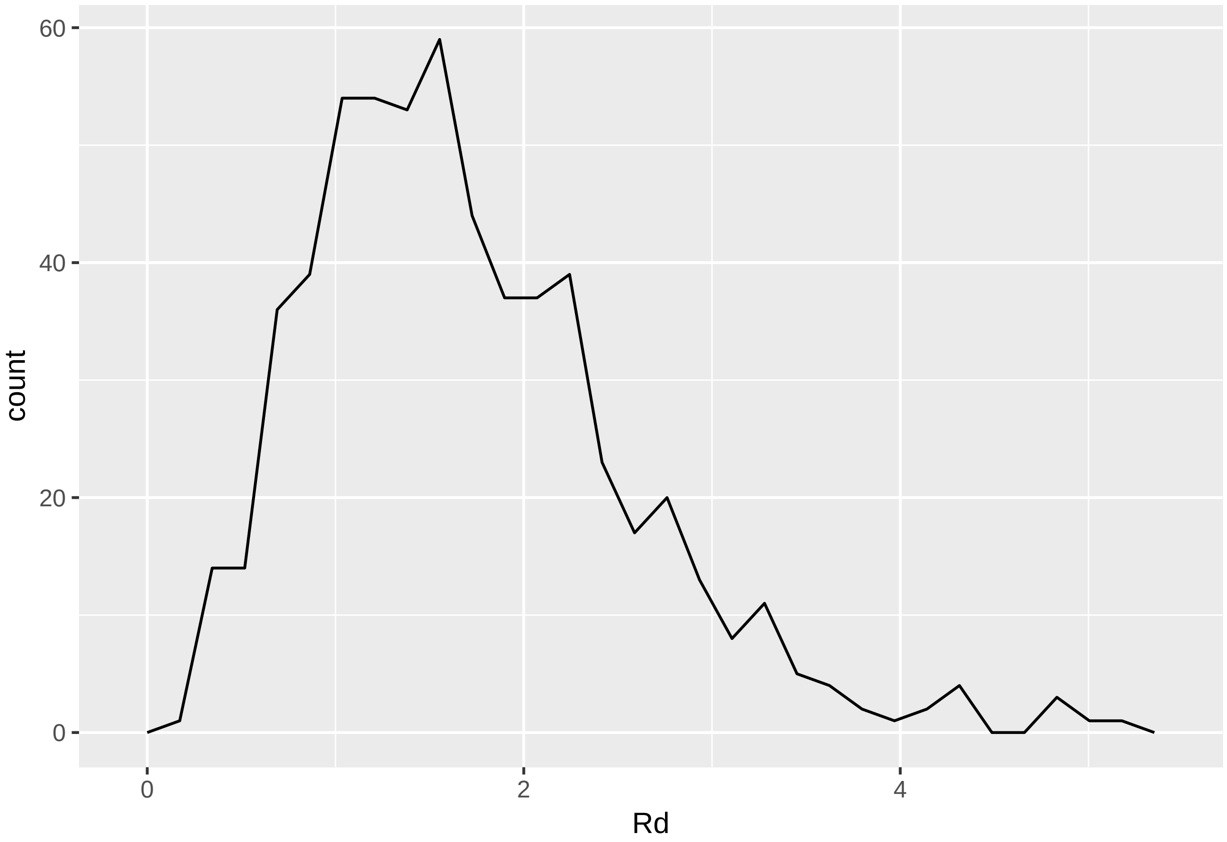
```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



```
# Plot Rd distribution
ggplot(data = LCE_global) +
  geom_freqpoly(aes(x = Rd))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

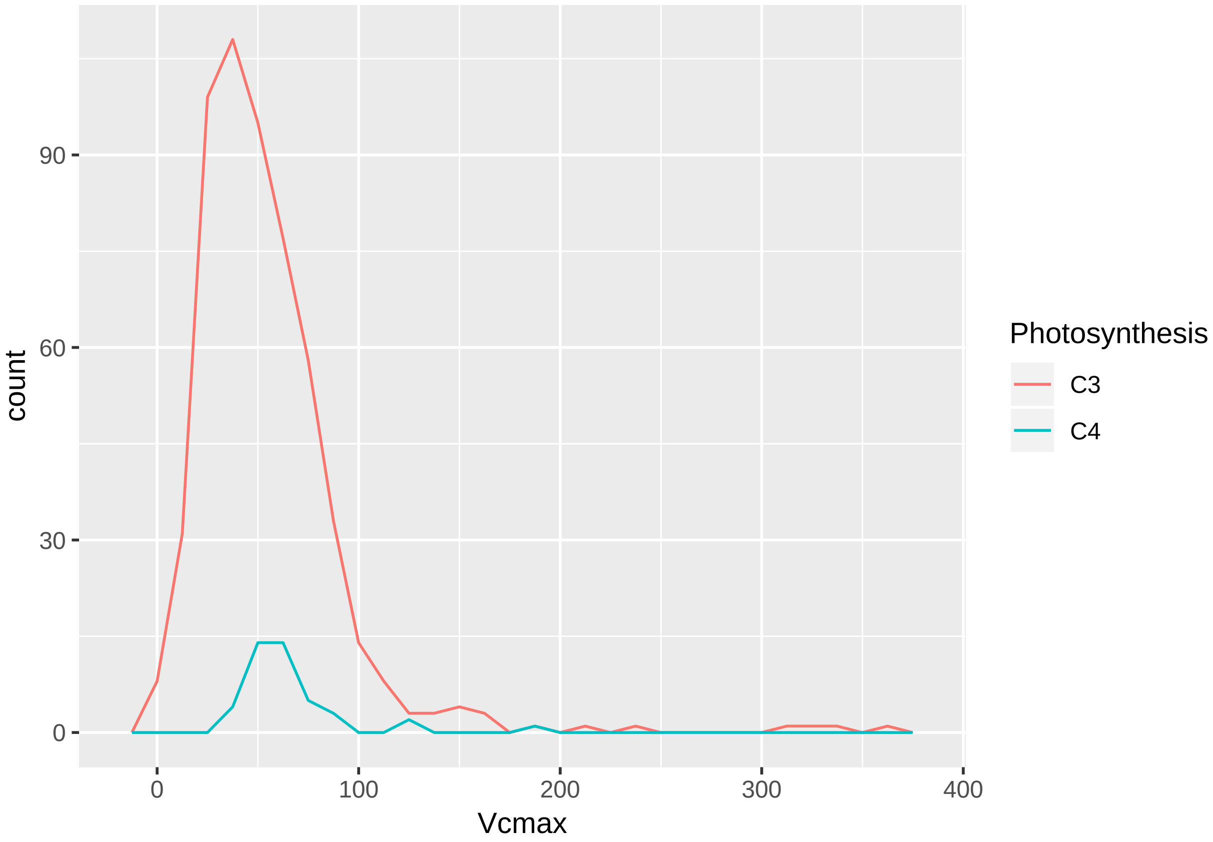


We can also use visual techniques to examine the covariation between a continuous and categorical variable. Let's start by examining the distribution for Vcmax within C3 and C4 plants:

```
# Plot distributions of Vcmax, colored by C3/C4 Photosynthesis
ggplot(data = LCE_global) +
  geom_freqpoly(mapping = aes(x = Vcmax, color = Photosynthesis))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

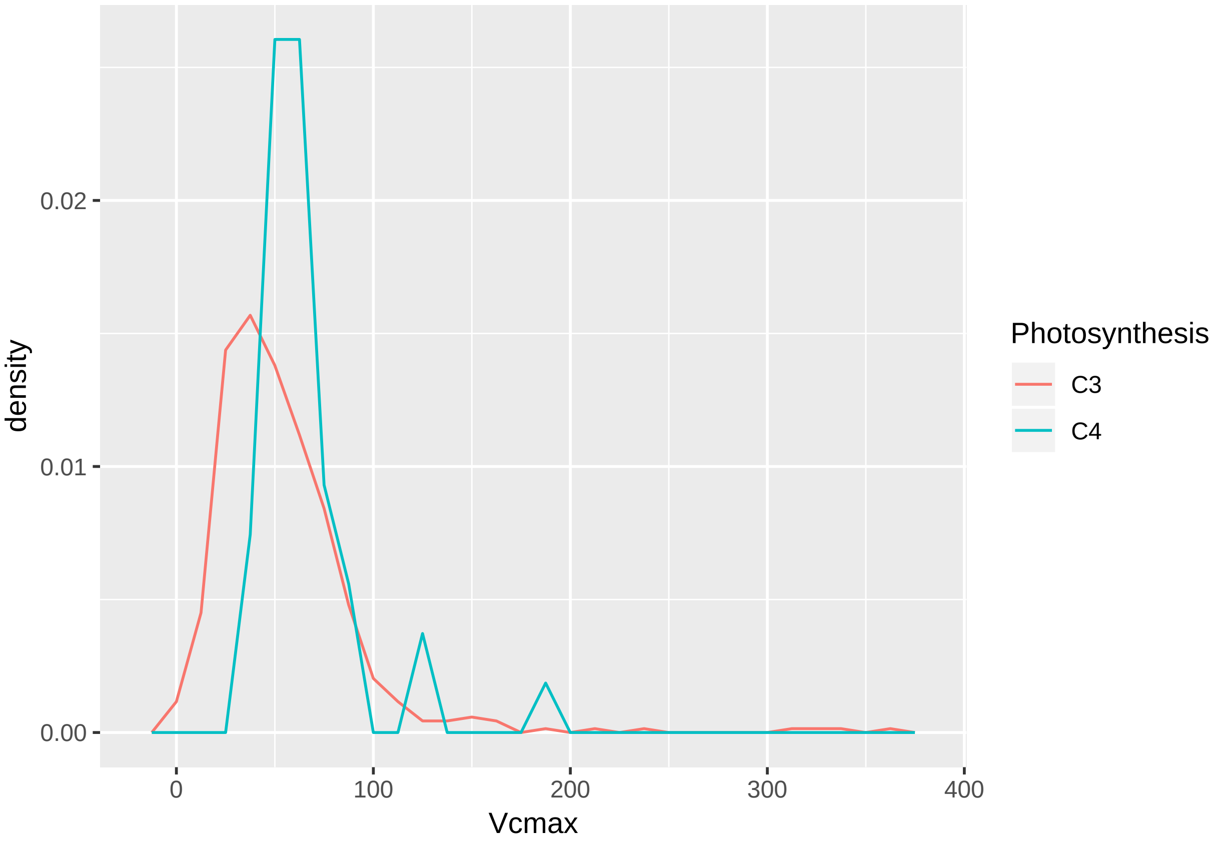
```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



```
# Alternate Plot: scaled by probability density (integrates to 1.0, rather than showing counts)
ggplot(data = LCE_global) +
  geom_freqpoly(mapping = aes(x = Vcmax, y = ..density.., color = Photosynthesis))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



Visualizing the distributions of the data will help to understand the range of values, show differences among categories, highlight interesting aspects of your data (groups, outliers, etc.), and eventually inform details of statistical tests that you might use to quantify differences among groups in your data.

Code Challenge 2: Examine the distributions for Vcmax separated by one of the other categorical variables in the data.

Code Challenge 3: Visualize the distributions for Vcmax separated by TWO of the categorical variables in the data.

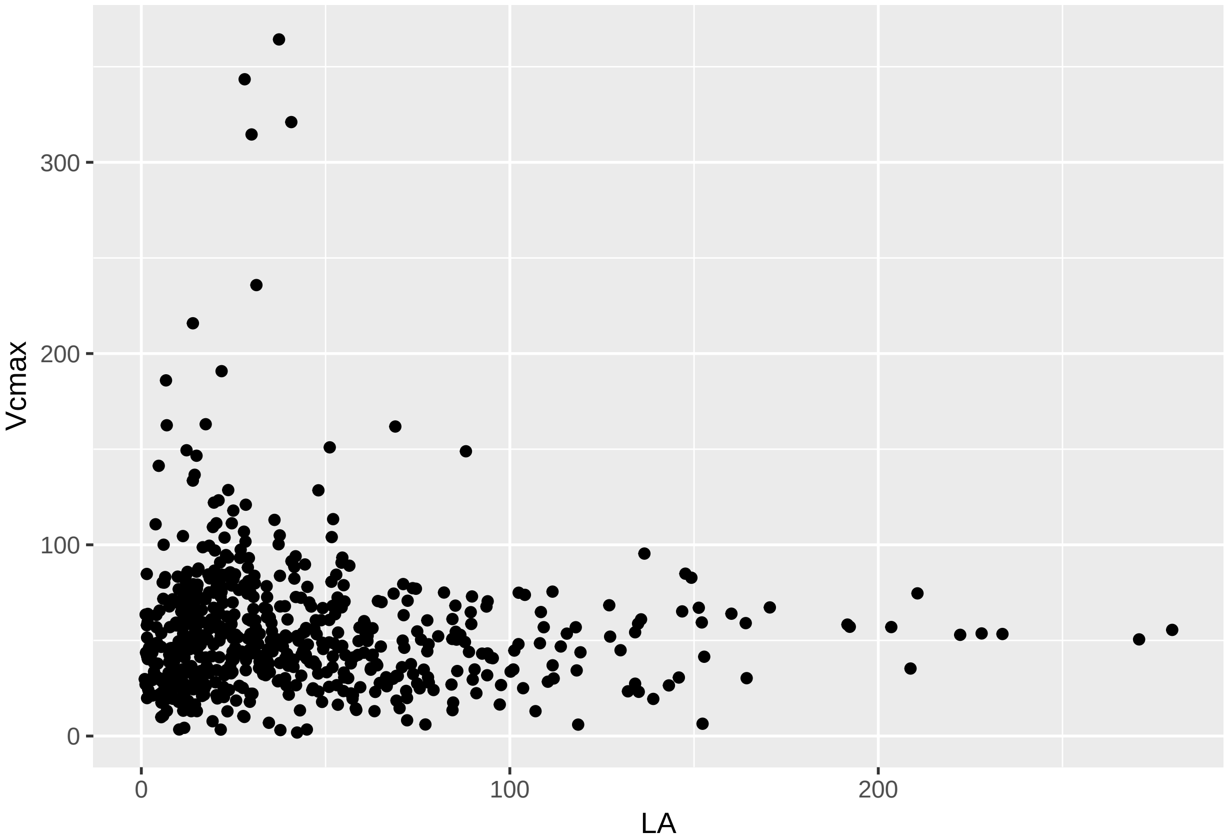
---

5. Visualizing covariation among two continuous variables

There are also several continuous variables for which we can examine the connection to the leaf carbon exchange variables within this dataset. Let's first look to see whether leaf mass (`LM` in the tibble) is connected to Vcmax:

```r
# Plot leaf mass (LM) against Vcmax
ggplot(data = LCE_global) +
  geom_point(mapping = aes(x = LA, y = Vcmax))
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```



Code Challenge 4: Visualize the covariation between two of the photosynthesis variables (Vcmax, Jmax, and Rd). Do they covary? If so, describe the pattern, and speculate about why they might covary.

6. Patterns and models

I think that the R for Data Science book does an excellent job outlining these four questions to ask yourself when you notice a pattern in your data:

1. Could this pattern be due to coincidence (i.e. random chance)?
2. How can you describe the relationship implied by the pattern?
3. How strong is the relationship implied by the pattern?
4. What other variables might affect the relationship?
5. Does the relationship change if you look at individual subgroups of the data?

Visualizing your data is arguably the most powerful way to look for patterns and to also convince others that these patterns exist. It's always helpful to start with visualization and exploratory data analysis to examine patterns in your data that address your research question. However, at some point you will need to do a statistical test with a statistical model to determine whether the patterns that you see are "significant" (this is a complex term that we'll only scratch the surface of within this class).

```r
# ANOVA: analysis of variance
# model differences in continuous variable against cateogorical groups
C3C4_Vcmax <- lm(Vcmax ~ Photosynthesis ,data = LCE_global)
summary(C3C4_Vcmax)
```

```
##
## Call:
## lm(formula = Vcmax ~ Photosynthesis, data = LCE_global)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.336 -22.392  -7.177  13.060 310.082
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         54.175      1.608  33.683   <2e-16 ***
## PhotosynthesisC4    10.323      5.978   1.727   0.0847 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.75 on 592 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.005012,   Adjusted R-squared:  0.003331
## F-statistic: 2.982 on 1 and 592 DF,  p-value: 0.08472
```

```r
# Tukey Pairwise Honestly Significant Differences (HSD) Test
TukeyHSD(aov(C3C4_Vcmax))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = C3C4_Vcmax)
##
## $Photosynthesis
##           diff       lwr      upr     p adj
## C4-C3 10.32266 -1.417599 22.06293 0.0847189
```

The output from this statitsical model and the Tukey HSD test show us that the mean Vcmax values for the two groups (C3 and C4 plants) are marginally significantly different ($p < 0.1$). The mean Vcmax for the C4 group is 10.323 (with standard error 5.978) larger than the C3 group.

One important qualification for these results is that whether a plant is C3 or C4 in our dataset might be confounded with other variables, and thus it might not represent a causal relationship. For example, if all of our C4 measurements were from a tropical site, this might mean that temperature, not photosynthesis type, was the causal mechanism for this pattern. Throughout exploratory data analysis and statistical testing, it's **essential** to keep reminding yourself about what the data mean and how different variables might be connected to each other.

For the purposes of this lab, a useful stratgey to avoid suprious or confounded results is to take a subset of the data (using Data Transformation tools from previous labs) to examine patterns at a smaller scale. For example, we could fit an ANOVA model for the Vcmax of C3 and C4 plants on data just from the tropical sites, or just from a particular site. Or, if we wanted to look at patterns across all the sites, we could subset to just examine C3 plants (or vice versa, C4 plants).

LAB REPORT INSTRUCTIONS:

- For your Lab 2 Report, you can investigate a question/hypothesis related to Vcmax, Jmax, or Rd that can be tested by the global leaf carbon exchange dataset. You can look at the entire dataset, or a subset (particular sites, particular species, etc.) to formulate/answer your research question.
- As you structure your data analysis to answer your question, produce an .R script pretending that you

are starting from scratch (i.e., don't assume that you have anything loaded from doing the lab exercise). The goal is to be able to hand someone your code and be able to have them re-run your analysis to see what you did and how - this is reproducible research!

- In addition to your .R script, for the Lab 2 Report you will turn in a text .pdf document no longer than 4 single-spaced pages in the format oulined within the Lab Report Guidelines.
- Your Lab 2 Report document must include at least one ggplot figure and one summary table, which counts toward the 4-page limit.