Lab 6: My Dataset

Enter a brief title/description for your data here

Your Name Here

Enter date here

General Information

This lab is due Wednesday, October 14th by 11:59 pm and is worth 10 points. You must upload your .rmd file and your knitted PDF to the assignment folder on Canvas.

Overview of the Independent Project

Each student will undertake an independent project during the semester. The primary objective is to have students statistically analyze a dataset of their own choosing. Students are encouraged to identify a real dataset that is interesting or meaningful to them, but the dataset must be approved by the instructor to ensure that it will meet the learning objectives of the course. Students, who are unable to find an appropriate dataset should consult with the instructor who will provide them with an appropriate dataset.

The objective is to conduct the analyses necessary to address the outlined hypothesis. The focus will be on understanding the data, developing a model appropriate to answer the question of interest, testing and assessing the model, and interpreting and displaying the results. This will include two assignments. Partway through the semester, students will be asked to turn in a document identifying their chosen dataset, briefly summarizing at least three scientific articles relevant to their topic, and preliminary plots (in R) of their response variable(s) and relationships with selected predictors. The final project will constitute a full report including Statistical Methods, Results, and Graphics in the style of a manuscript being prepared for submission. This means students will have to critically address which results to report, how to report them, and which figures and tables to present. This report will be graded on the basis of statistical practice, quality of reporting, support for interpretation, and quality of figures.

What is an Appropriate Dataset?

Philosophy

The most important thing is that the dataset addresses a biological question in which you are interested. Second, the point is NOT to recreate an existing analysis, but to instead discover something new entirely on your own. So if you choose to use a dataset from a previously published paper, then you will need to use these data in a fresh new way. The whole point of this exercise is to move beyond 'canned labs' where the results are already known (by the professor at least) and to use the skills that we have learned this semester to discover something entirely new! In previous courses, I have had some students who have published the results of their independent project in the peer reviewed literature. This is not my expectation for all projects, but I hope that you will strive to complete novel and impactful work!

Some specific guidelines

- At least 50 records (or observations or rows)
- A continuous response variable or a discrete response variable (0 or 1; counts; etc)
- At least 5 covariates/predictors (continuous and categorical)

Where should I look for a dataset?

If you do not already have data from an undergraduate or graduate thesis project then talk to your supervisor or one of your instructors who does research that interests you and ask whether they have data that you might use. Remember that the goal is to do something new and not to repeat a previous analysis. You can also look for datasets online (e.g. Dryad, FigShare, Ecological Archives, etc). If you still cannot find a dataset then come talk to me. I can provide you with one, but it will be easiest for you if the data already mean something to you and contains variables that you are interested in. I will also post some possible datasets from the course in previous years as well as some explanations of their variables on Canvas.

This Assignment

The goal of this assignment is to make sure that everyone has both a dataset to use for their final project.

Part A (6 points)

Provide a brief description of the biology behind your dataset. Who collected the data and for what purpose? What level of biological organization was sampled as the unit of replication (e.g. individual, population, community, ecosystem, etc.)? Were these units sub-sampled? Were the data collected as part of an experiment or were they observational? How were replicate samples collected (i.e. randomly, systematically, opportunistically)? You do not need to have a research question or hypothesis at this stage.

The data I chose to analyze comes from a study conducted from 2006 to 2009 by Elits et al. Their goal was to investigate the effects of clonal plant species on total species richness across nutrient treatments that were applied in varying degrees of spatial heterogeneity. The study was experimental. Samples were collected individually (species count for a spatially heterogeneous nutrient plot). Plants were sowed from seed in each given plot, and these plants were selected from a seed bank of species that were common to the study cite (no indication as to how these species were selected). This methodology applies to both clonal and non-clonal species used in the experiment. The basic biology behind their study was that clonal plants decrease species richness when communities (and the soil nutrients within) are spatially heterogeneous. Clonal plant species are able to extend beyond their immediate community and infiltrate adjacent ones, usurping resources, growing in biomass, and thus decreasing the species count in each community they exist in through competitive exclusion. In theory, this effect should be more noticeable when communities are close together small and less so when communities are larger, and thus further apart. By homogonizing fertilizer treatments across these plots, they hoped to learn more about this effect

Part B (4 points)

In the code box below, import your dataset into R. First make sure that your dataset is flat and 2-dimensional. Each row should correspond to a replicate or a sub-replicate and each column should correspond to a variable. There should be only one data file. You might find it easiest to make some changes to your variable names at this point so that they are relatively simple but informative for import into R.

```
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------- tidyverse 1
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.1
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ---------------------------------------------------------------------- tidyverse_conflic
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
richnessDat <-read_excel("~/Repos/School/BioStats/FinalProject/finalProjData.xlsx",
                         sheet = "Species Richness")
biomassDat <- read_excel("~/Repos/School/BioStats/FinalProject/finalProjData.xlsx",
                         sheet = "Biomass")
#'Removing years in which biomass measurements were not taken from richness data
richnessDat <- richnessDat[-which(richnessDat$Year == "2007"),]
richnessDat <- richnessDat[-which(richnessDat$Year == "2009"),]
richnessDat <- richnessDat[-which(richnessDat$Year == "2005"),]

#'Combining data
richnessDat$TotalBMS <- biomassDat$TotalBMS
richnessDat$NCBMS <- biomassDat$NCBMS
richnessDat$ClonBMS <- biomassDat$ClonBMS

totalDat <- richnessDat
```

Export or save this dataset from its current format to a comma-separated format (filename.csv). Here is some example code to export the object 'filename' as a csv file called "output_filename.csv":

```
# write.csv(filename, "output_filename.csv")
write.csv(totalDat, "Richness_and_Biomass_Lab6.csv")
```

The csv file will be output to your working directory.

In these instructions I will provide examples such as "filename" above. Please name your files and variables in a way that makes most sense to you. My names are just examples.

The point of exporting the csv file is because I will need a csv copy of your dataset in order to check and trouble-shoot your analyses. Please make sure I have a copy of your up-to-date datafile and do not make any changes to it without either including these in your code (here and in your final project submission) or sending me an updated (modified) file.

Use the summary command to provide descriptive statistics for the variables in your data file.

```
summary(totalDat)
```

```
##       Year          Block       Community         FertilizerTreatment
##  Min.   :2006   Min.   :1.00   Length:128         Length:128
##  1st Qu.:2006   1st Qu.:2.75   Class :character   Class :character
##  Median :2007   Median :4.50   Mode  :character   Mode  :character
##  Mean   :2007   Mean   :4.50
##  3rd Qu.:2008   3rd Qu.:6.25
##  Max.   :2008   Max.   :8.00
##   TotalSppNum       NCSppNum       ClonSppNum     TotalBMS
##  Min.   :12.00   Min.   :11.00   Min.   :0.0   Min.   : 80.17
```

```
##  1st Qu.:22.00   1st Qu.:19.00   1st Qu.:0.0   1st Qu.:129.67
##  Median :25.00   Median :23.00   Median :1.5   Median :178.90
##  Mean   :25.02   Mean   :22.52   Mean   :2.5   Mean   :211.25
##  3rd Qu.:28.00   3rd Qu.:25.00   3rd Qu.:5.0   3rd Qu.:285.47
##  Max.   :38.00   Max.   :34.00   Max.   :8.0   Max.   :479.12
##      NCBMS           ClonBMS
##  Min.   : 11.05   Min.   :  0.00
##  1st Qu.: 93.51   1st Qu.:  0.00
##  Median :127.18   Median :  1.05
##  Mean   :145.16   Mean   : 66.09
##  3rd Qu.:177.91   3rd Qu.: 88.04
##  Max.   :398.57   Max.   :373.05
```

Indicate which variables are dependent (i.e. response) variables (hopefully only one or a few) and which variables are independent (i.e. predictor) variables (hopefully > 5 as indicated in the first lecture). Indicate which variables, if any, were experimentally applied. For variables that are coded (e.g. Treatment = 1, 2, 3) be sure to indicate what each of these codes represents. Provide units for variables as necessary.

Here is an example from a recent file that I put together:

Variables and Their Explanations

id - A unique identifier for each squirrel

litter_id - A unique indentifier for each litter

Grid - One of two long-term control study areas (KL = Kloo; SU = Sulphur)

Sex - M or F; there is one squirrel of unknown sex

YearOfBirth - birth year

YearOfDeath - year in which squirrel was last sighted or captured alive. This might not be a good measure of the year in which the squirrel died and the timing within the year could be quite important. So I have also included dateE below, which should work better than dyear.

dateE - Date on which the squirrel was last seen or captured alive (yyyy-mm-dd). Squirrel was assumed dead after this point.

LBS - total number of offspring born to each squirrel born in the study.

LRS - total number of offspring recruited for each recruited squirrel. In our study recruitment is determined based on survival to 200 days of age. Because of our seasonal data collection, this requires survival to the following spring when squirrels could have bred for the first time. Squirrels that did not survive to 200 days of age have LRS = NA.

birth_conest - Cone index for the cone crop in the autumn of the year in which the squirrel was born.

birth_conestm1 - Cone index for the cone crop in the autumn prior to the year in which the squirrel was born.

birth_density - spring population density (squirrels/ha.) for the spring in which the squirrel was born.

mast - was the squirrel born in a mast year or not? (y or n).

Year - time point indicating when data was gathered

Block - the replicated block from which the data originated

Community - one of two experimentally created communities, non-clonal only where clonals were prevented from establishing, and mixed where both non-clonal and clonal species were present

Fertilizer Treatment - One of four fertilizer regimes: CO = control with no fertilizer added; UN = uniform addition of 20g N m-2 yr-1; LP = large patchy plots each plot containing four fertilizer levels, zero 0gN m-2 yr-1, low 13.4 gN m-2 yr-1, medium 26.7 gN m-2 yr-1, high 40.2 gN m-2 yr-1 each at 1.5m x 1.5m with the plot total summing to the same N added per year as UN; SP = small patchy with the same total N added as the UN and the same four levels of addition as LP distributed as 144 pathces of 25cm x 25cm with each fertilizer level equally represente

TotalSppNum total unique number of plant species counted within a plot

NCSppNum total number of non-clonal plant species within a plot contributing to the total species richness

ClonSppNum - number of clonal species within a plot contributing to the total species richness

TotalBMS - total biomass (g m-2) sampled at the peak of standing biomass in mid-summer from 12 randomly selected 25cm x 25 cm grid cells from 144 cells per plot

NCBMS - all non-clonal species biomass (g m-2) contributing to the total sampled biomass

ClonBMS - all rhizomatous species biomass (g m-2) contributing to the total sampled biomass

Summarizing the data

Indicate the number of rows in your data file by using:

```
length(totalDat$Year)
```

```
## [1] 128
```

If you have factors (i.e. categorical variables) then make sure that they are included as factors. If not, use the factor command to change them.

```
totalDat$Community <- as.factor(totalDat$Community)
totalDat$FertilizerTreatment <- as.factor(totalDat$FertilizerTreatment)
```

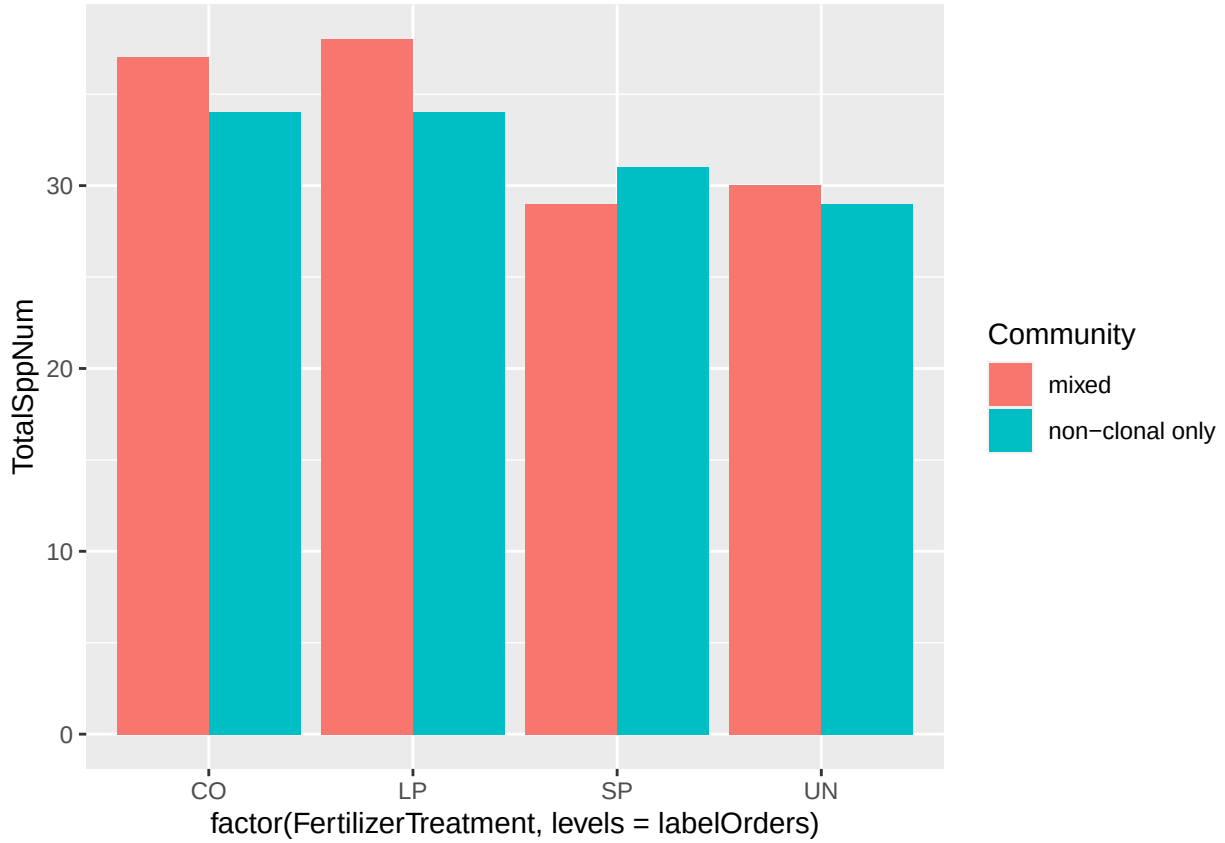
Any individual, site or plot identifiers should be factors. Recall the summary command. Indicate what the variables and their levels mean as above.

Some Basic Plots

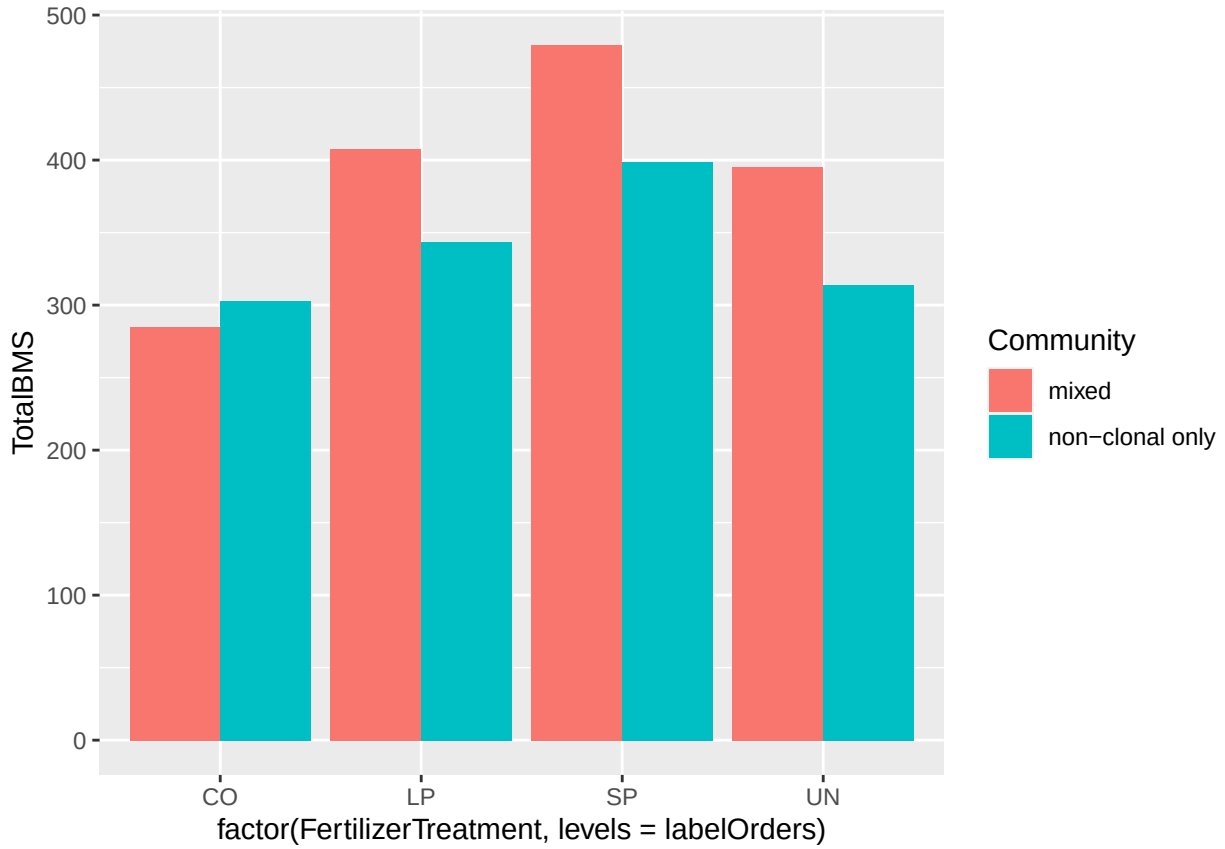
Provide 2 preliminary plots (in R) of the response variable(s) and/or its relationship with selected predictors. What kind of plot will depend on the type of relationship. These could include a histogram of the distribution (raw or transformed), bivariate scatterplots with numerical predictors, or a stripchart or even mosaic plot for categorical predictors.

```
#'Label orders for vis
labelOrders <- c("CO", "LP", "SP", "UN")

#Total SPP num by fertilizer treatment
ggplot(totalDat, aes(x = factor(FertilizerTreatment, levels = labelOrders), TotalSppNum))+
  geom_bar(stat = "identity", aes(fill = Community), position = "dodge")
```



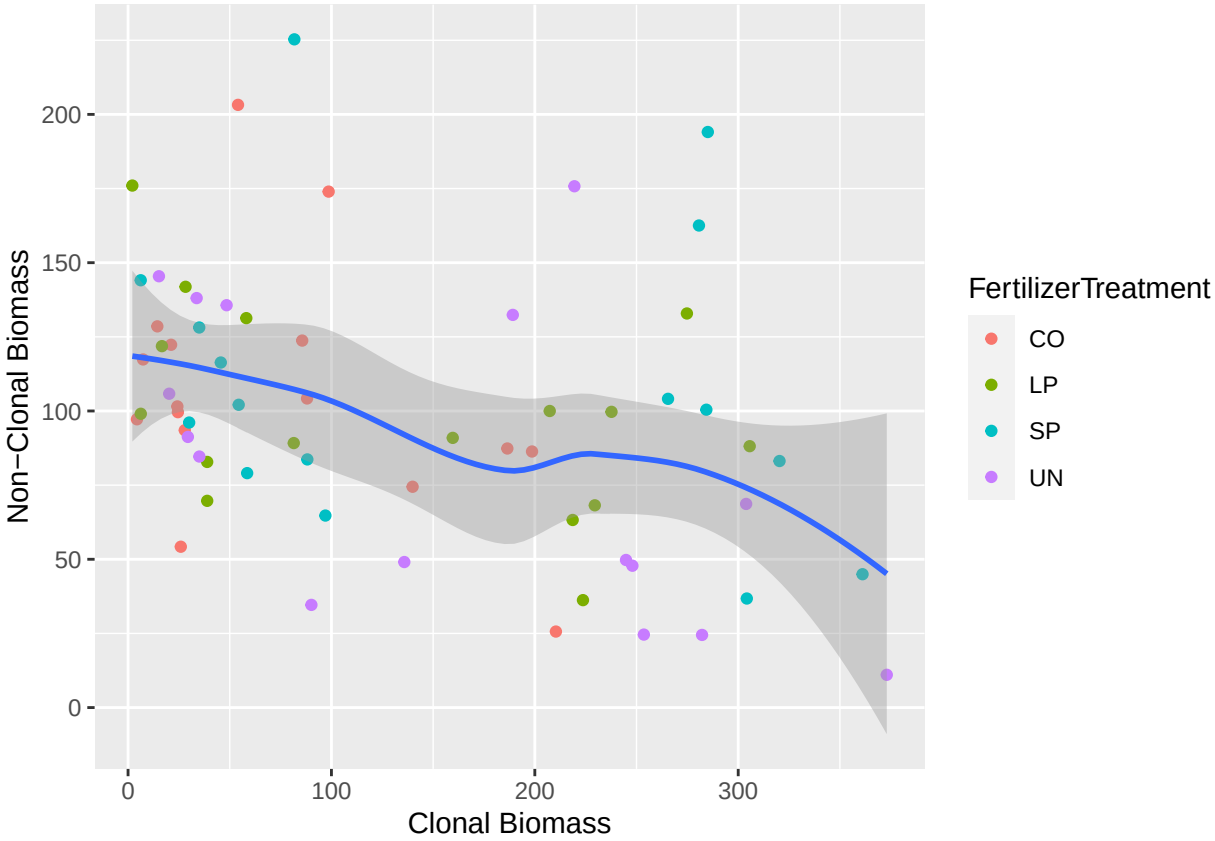

```
#Total Biomass by fert. treatment
ggplot(totalDat, aes(x = factor(FertilizerTreatment, levels = labelOrders), TotalBMS))+
  geom_bar(stat = "identity", aes(fill = Community), position = "dodge")
```




```
NCDat <- totalDat %>%
  filter(Community == "mixed")

#'Clonal Biomass as a predictor of NonClonal Biomass
ggplot(NCDat, aes(ClonBMS, NCBMS)) +
  geom_point(aes(col = FertilizerTreatment)) +
  geom_smooth(method = "loess") +
  labs(x = "Clonal Biomass", y = "Non-Clonal Biomass")
```

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



Steps Moving Forward: I want to address the following questions: 1) Does non-clonal plant biomass differ within each fertilizer plot by community? 2) Does clonal biomass have an effect on non-clonal biomass?

Save your R Markdown file and knit a PDF of the file and output.

What to submit

- A copy of your csv file
- A knitted PDF of your assignment
- A copy of your Rmd file