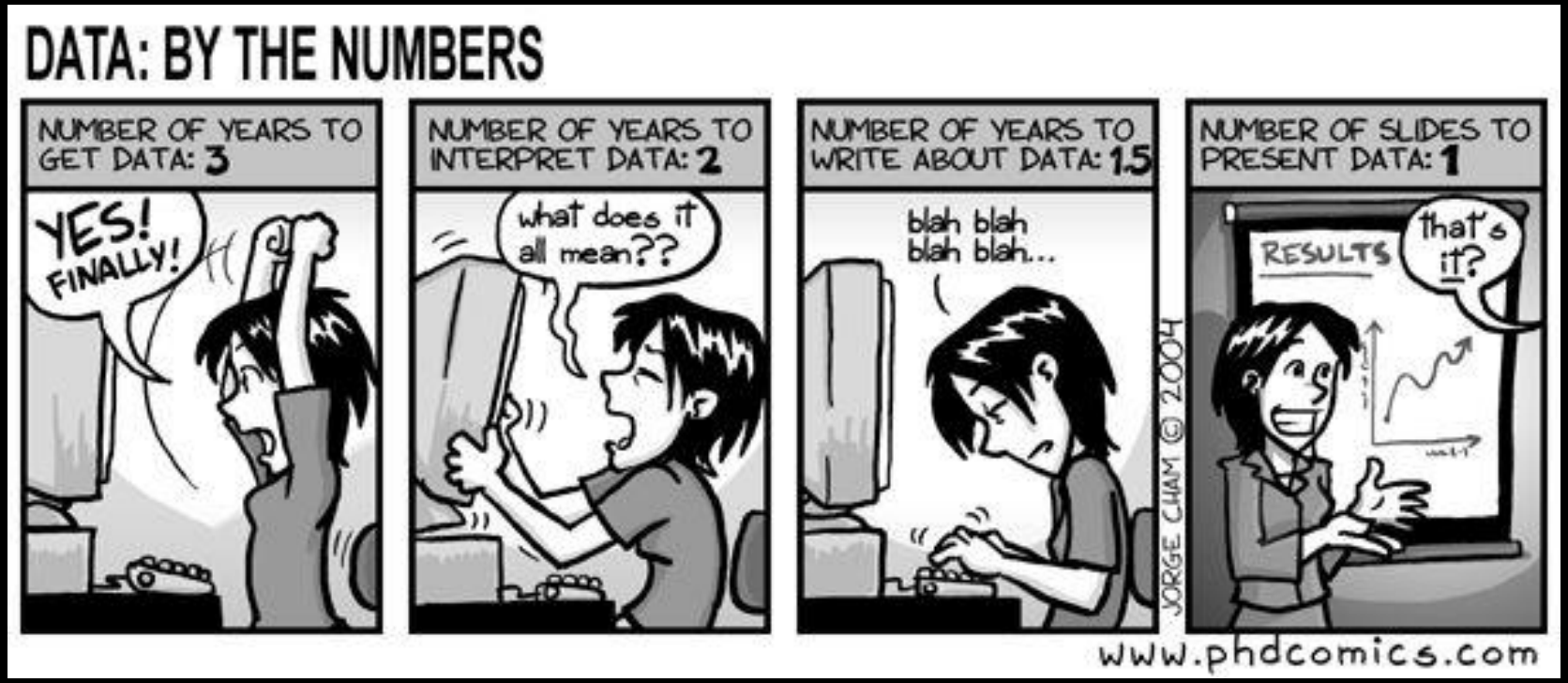
Welcome to BIOL 599: Ecological Data Analysis

Dr. Xanthe Walker; xanthe.walker@nau.edu



As you come in, please **make a name tent**, and **write in BIG font**

Outline for today

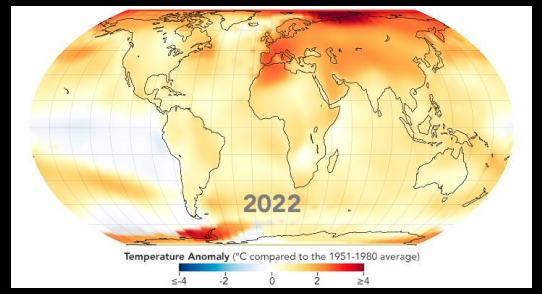
- Meet your Instructor
- About the course
- Why we use R
- Organizing data for analysis in R
- Wrap-up
  - To Do List before Thursday's workshop
  - Pre-course survey
  - Sign up for Discussion (2 students per presentations – on canvas)

Who am I?

- Xanthe Walker (she/her):
  - Assistant Professor in:
    - ECOSS (Center for Ecosystem Science and Society)
    - Biological Sciences
    - SICCS (School of Informatics, Computing, and Cyber Systems)
  - At NAU for 10 years

What do I research?

↑[CO_2]



↑ Temperature

↑ Fire

Δ Ecosystems

?

Carbon source
Accelerating
Positive feedback

Carbon cycle

Carbon sink
Mitigating
Negative feedback

Who am I?

Who are you?

- Name and Program
- Research Interests
- Rose and thorn (best/worst) of your summer

Tell me more about you……

Pre-course survey posted on Canvas

- Who are you academically?
- Who are you outside of your thesis?
- Familiarity with R?
- Familiarity with stats theory?
- What are some potential types of data you want to collect for your thesis?
- Types of analysis you are interested in learning in R?

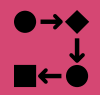
My stats experience

I'm not a statistician, but I enjoy using stats to find patterns that our eyes can't see.

I learned on Base R before Rstudio existed

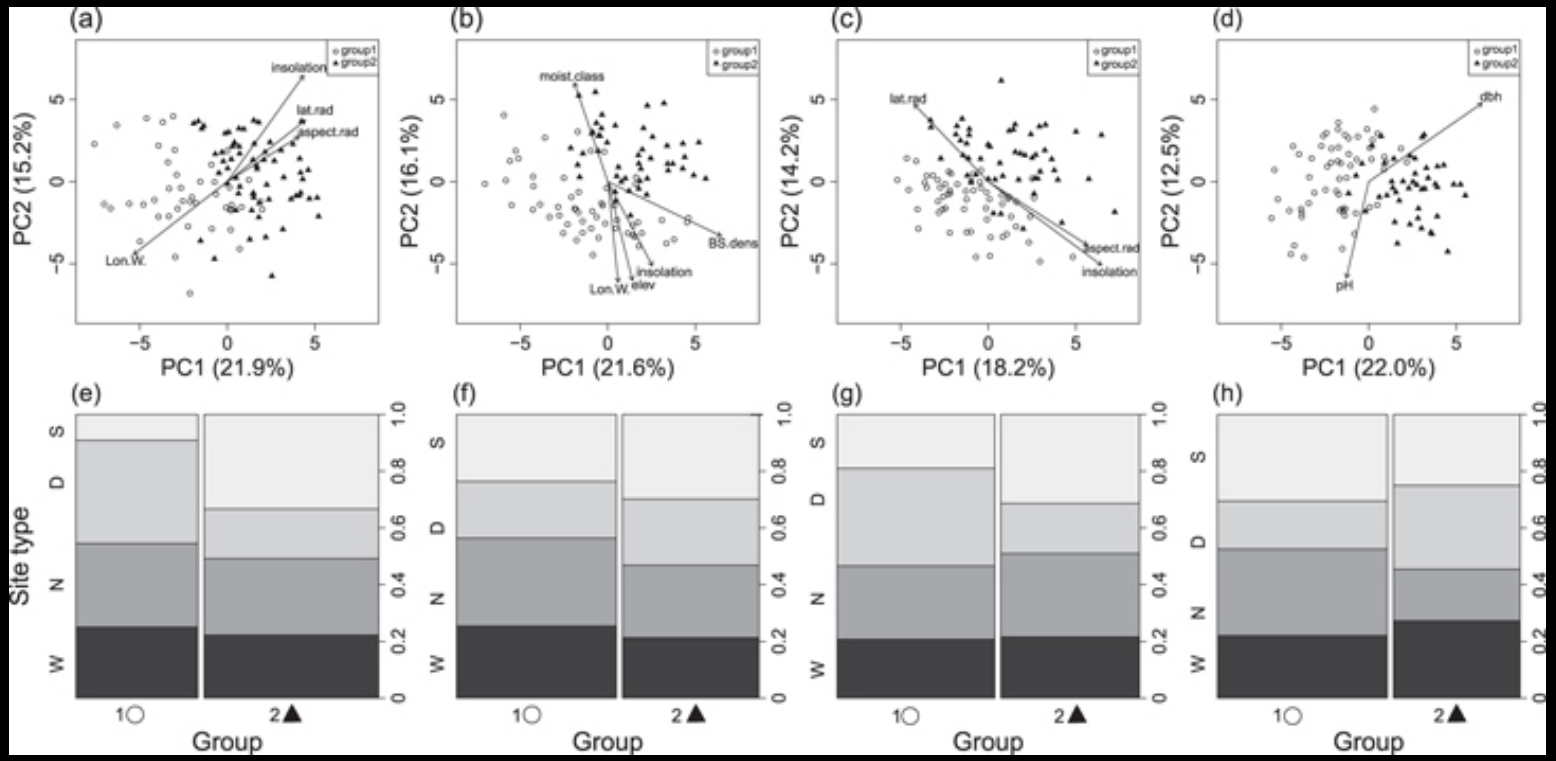
I am not an R expert. I have a head start on most of you. Maybe I will hold this edge until the end of term.

I am self-taught.

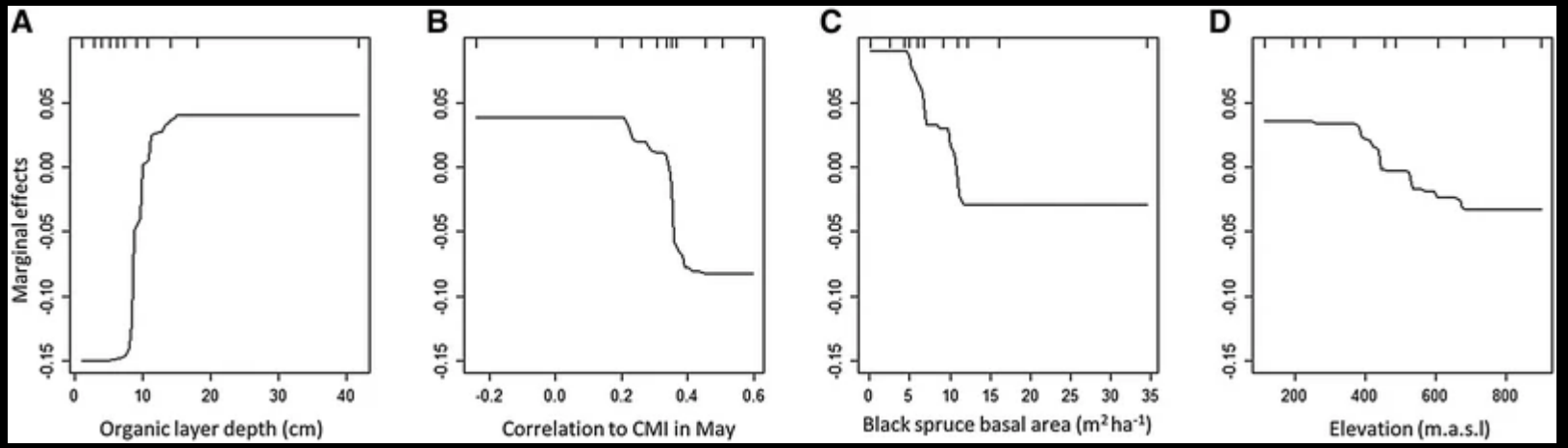


You will discover things that I don't know about. Please share these with me.

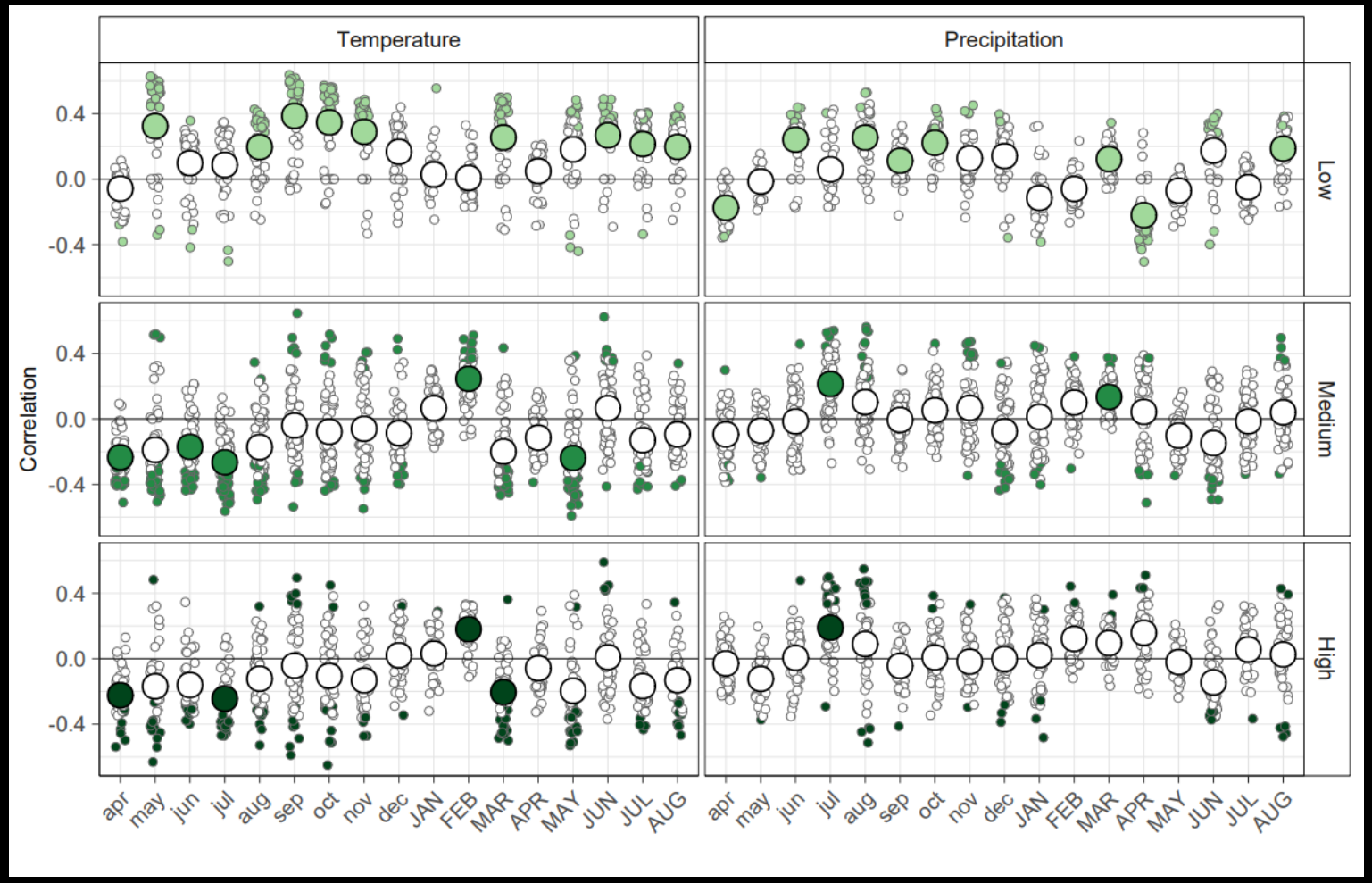
What type of analyses have I published?



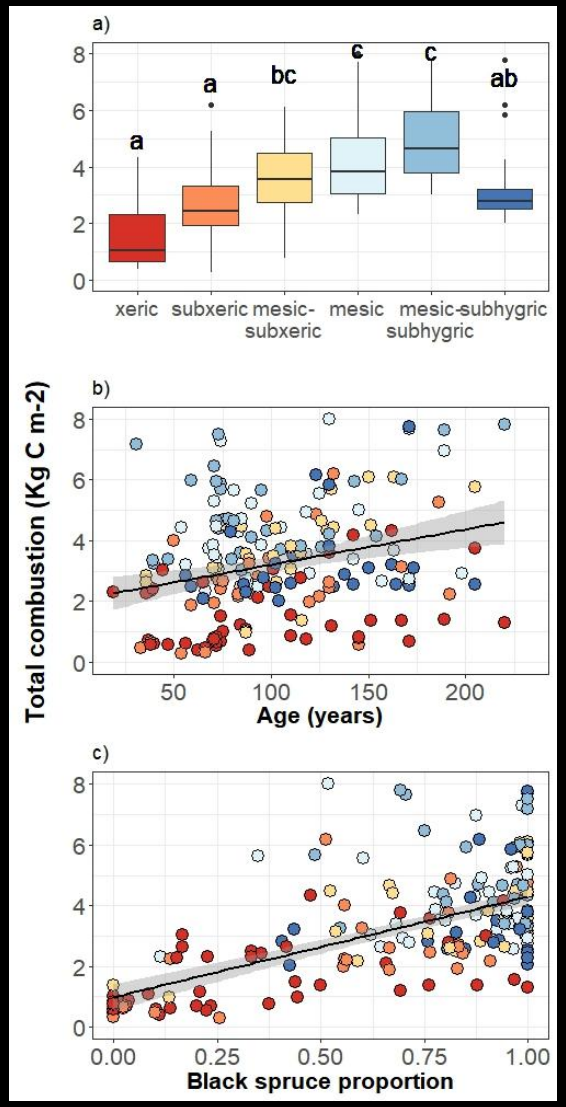
What type of analyses have I published?



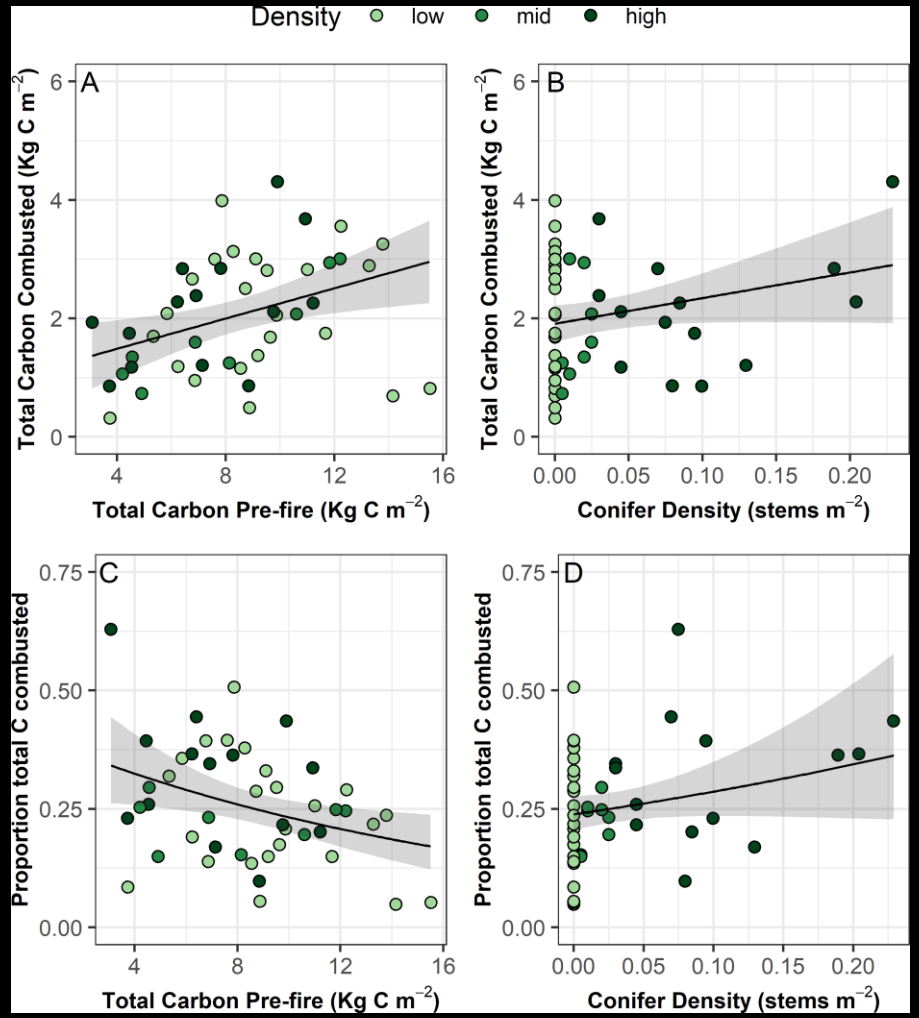
What type of analyses have I published?



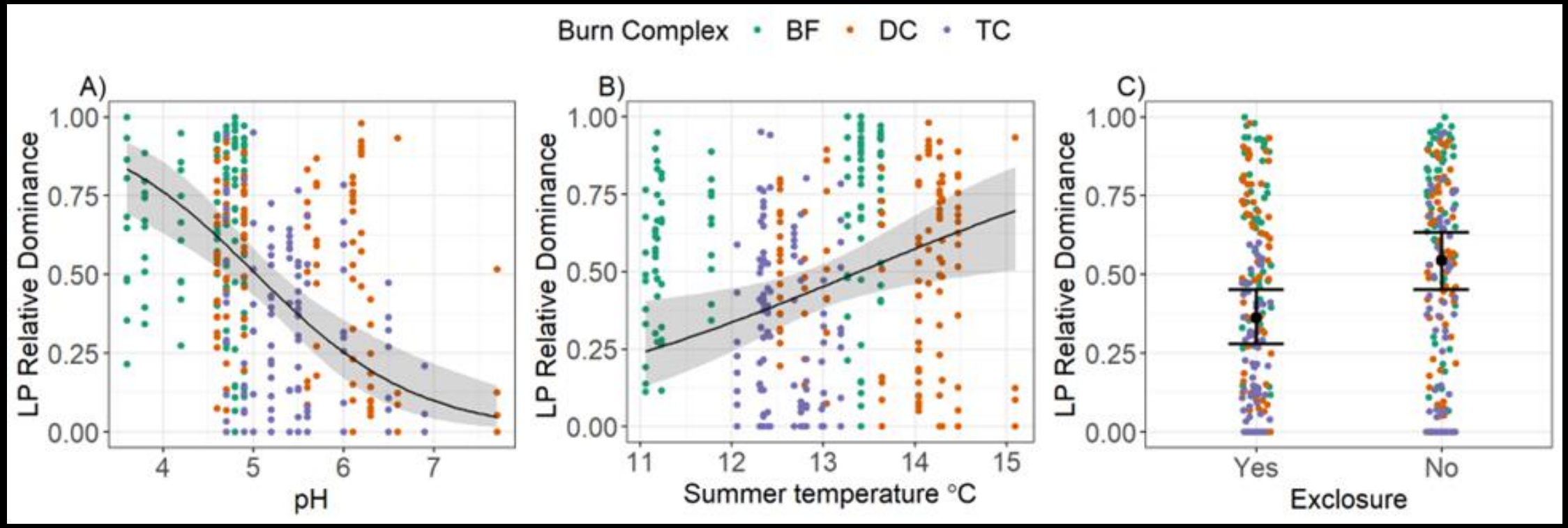
What type of analyses have I published?



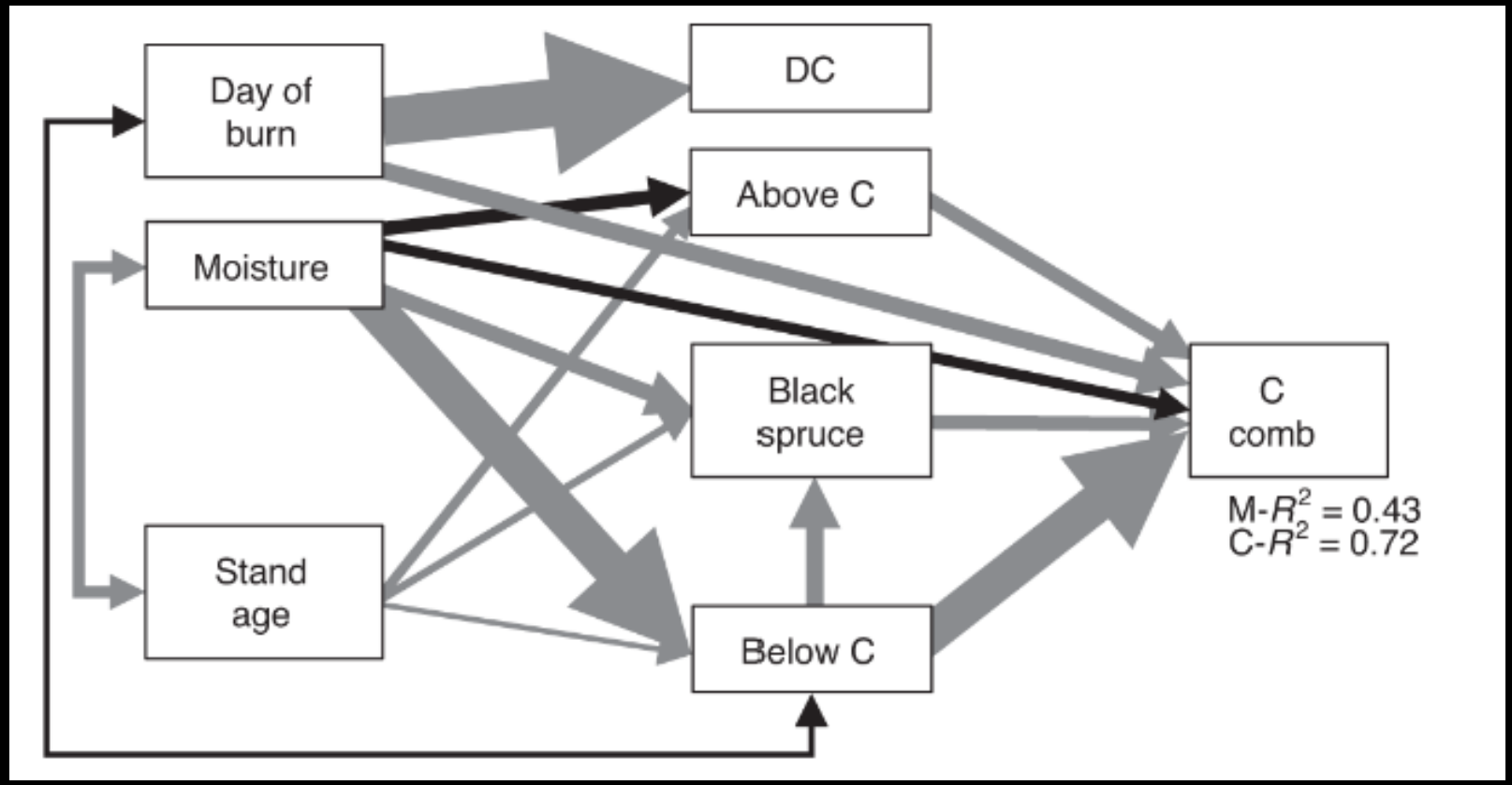
What type of analyses have I published?



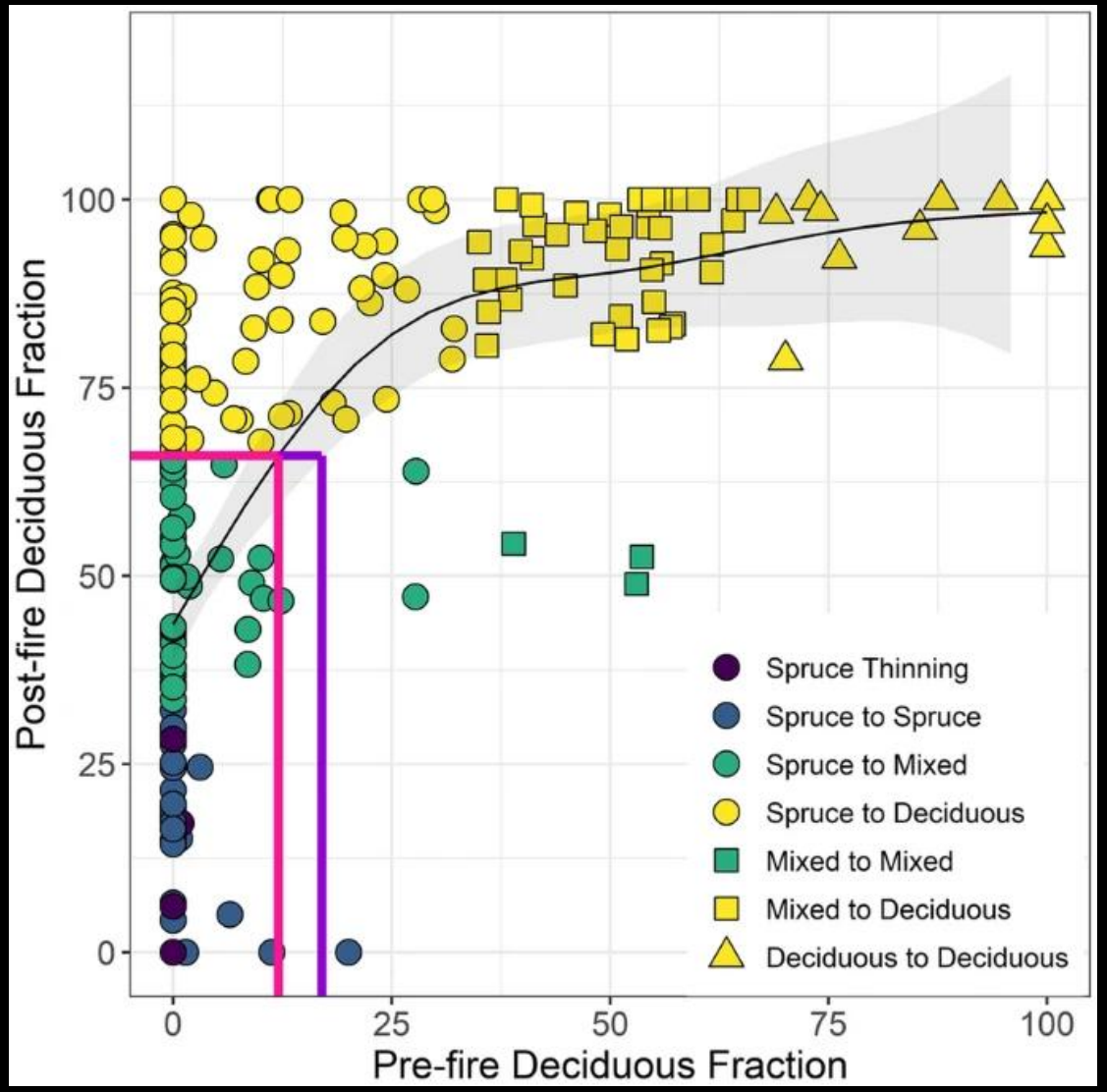
What type of analyses have I published?



What type of analyses have I published?



What type of analyses have I published?



Discussion

How do you feel about statistics and coding?

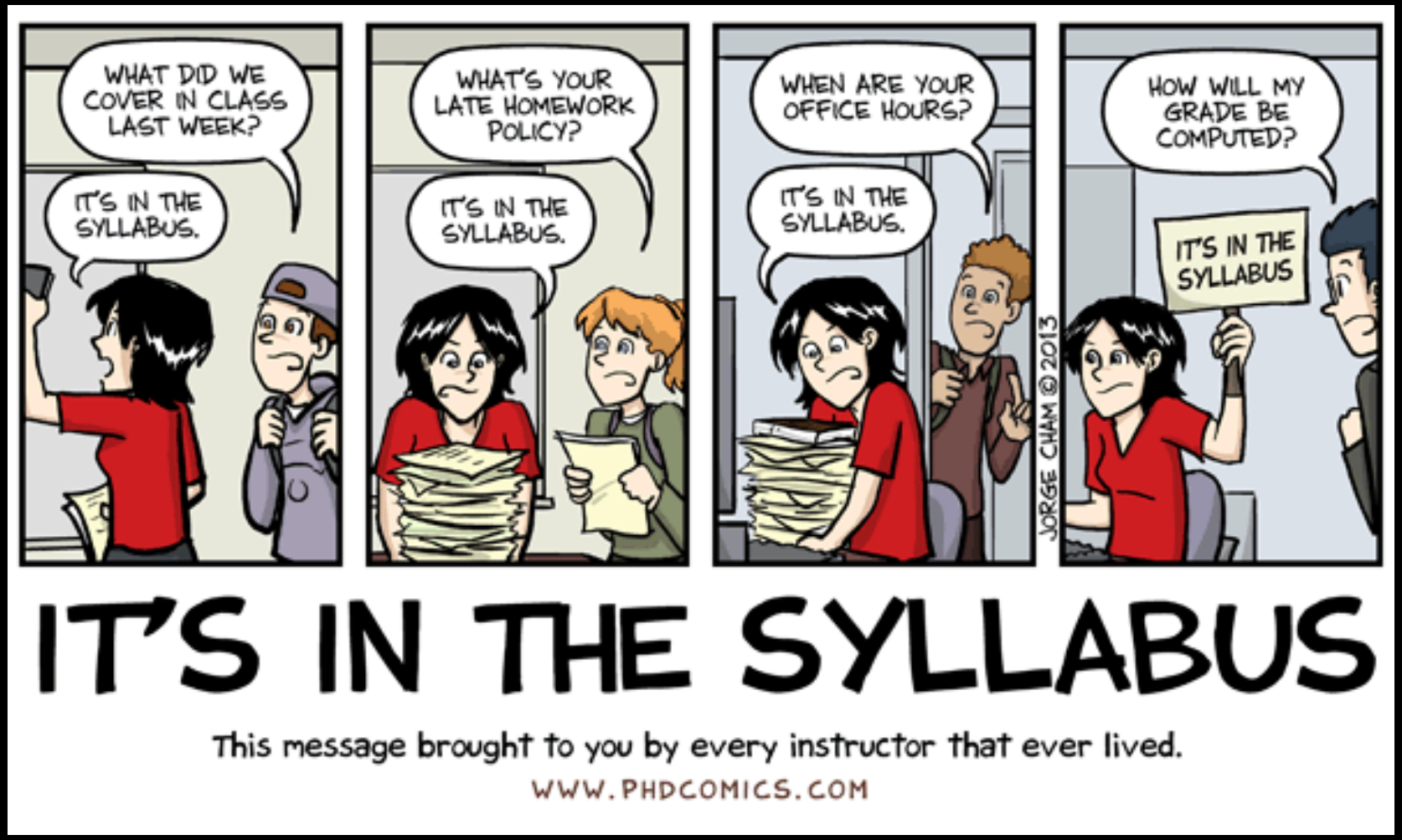
Outline for today

- Meet your Instructor
- About the course
- Why we use R
- Organizing data for analysis in R
- Wrap-up
  - To Do List before Thursday's workshop
  - Pre-course survey
  - Sign up for Discussion (2 students per presentations – on canvas)

About the course

- Course organization
- Schedule
- Grades
- Assignments
- Policy Statements

Please see Canvas syllabus for details
Today is just highlights



Course Background

This is a “second” course in data analysis, to take you beyond the most basic, introductory level, which I’m assuming you have already done

Help me improve it: complete the anonymous feedback forms

Learning Objectives

1. Prepare you for research by reviewing the basic principles for designing good studies, gathering and organizing data, and properly analyzing those data.
2. Develop proficiency in the programming language R to manipulate, summarize, analyze, and interpret data.
3. Choose appropriate analysis techniques for a variety of data types and formats.
4. Understand regression-type analysis ranging from simple linear regression to more complex generalized linear mixed effects models, and how to apply them.
5. Create publication-ready graphics and learn what to write in a paper.

Course format

Lecture (Tues)

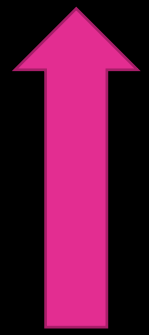
- Overview and background of the topic/method
- Work through an example with R code
- PPT will be posted on Canvas Monday prior to class

Discussion & Workshops (Thurs)

9:35-9:50
Student-led Paper Discussion

9:50-10:50
Self-paced workshop

Turn in completed Exercises – graded for completion (following Monday)



Start the workshop on Wednesday

Course Schedule

Week	Topic	Assignments	Reading
1	Introduction to course, R, Rstudio, Rmarkdown, Data visualization		
2	Exploratory data analysis		Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution, 1(1), 3-14. https://doi.org/10.1111/j.2041-210X.2009.00001.x
3	Linear Regresion- bivariate and multiple – limitations in ecology	Assignment #1	Boldina, I., & Beninger, P. G. (2016). Strengthening statistical usage in marine ecology: Linear regression. ICES Journal of Marine Science, 73(6), 1659-1664. https://doi.org/10.1093/icesjms/fsw066
4	GLS: Generalized Least Squares		Cleasby, I.R., Nakagawa, S. Neglected biological patterns in the residuals. Behav Ecol Sociobiol 65, 2361–2372 (2011). https://doi.org/10.1007/s00265-011-1254-7
5	Mixed effects models (MEM)		Harrison, X. A., et al. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ, 6, e4794. https://doi.org/10.7717/peerj.4794
6	Model Selection and Collinearity	Assignment #2	Tredennick, A. T., G. Hooker, S. P. Ellner, and P. B. Adler. 2021. A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology 102(6):e03336. 10.1002/ecy.3336
7	Violation of Independence; spatial and temporal autocorrelation		F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., R. Peres-Neto, P., Reineking, B., Schröder, B., M. Schurr, F. and Wilson, R. (2007), Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography, 30: 609-628. https://doi.org/10.1111/j.2007.0906-7590.05171.x
8	GLM: Generalized Linear Models		O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. Methods in Ecology and Evolution, 1(2), 118-122. https://doi.org/10.1111/j.2041-210X.2010.00021.x
9	Machine Learning (Jeremy Forsythe)		
10	Zero Inflated Models (ZIP & ZAP)		Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., & Castells, E. (2019). What does a zero mean? Understanding false, random and structural zeros in ecology. Methods in Ecology and Evolution, 10(7), 949-959. https://doi.org/10.1111/2041-210X.13185
11	GLMM: Generalized linear mixed models		Austin PC, Kapral MK, Vyas MV, Fang J, Yu AYX. Using Multilevel Models and Generalized Estimating Equation Models to Account for Clustering in Neurology Clinical Research. Neurology. 2024 Nov 12;103(9):e209947. doi: 10.1212/WNL.0000000000209947. Epub 2024 Oct 11. PMID: 39393031; PMCID: PMC11469681.
12	What to write in a paper	Assignment #3	Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2016). A protocol for conducting and presenting results of regression modelling in ecology. Methods in Ecology and Evolution, 7(6), 636-645
13	Review and Final project work		
14	Final project work		
15	Final project work	Final Project	

Grading Scheme

1. Weekly R markdown files (20%) from in-class R workshops (week # 1 is optional)
2. Lead a discussion of an assigned paper (10%).
3. Three minor Assignments (40%).
4. Final Project (20%).
5. Attendance and participation (10%).

Grading rubrics are posted – use these rubrics to complete your assignments!

Student-led paper discussion (Thurs)

- Two students will lead each paper discussion
- All students read the paper and participate in the Discussion
- **Presentation of paper (~5-10 min)**
  - Present a summary of paper
  - Explain key points, figures, additional opinions
- **Discussion of paper (~5-10 min)**
  - Moderate the discussion
  - Whole class participation

Each person please sign up for 1 Discussion spot on the Google Doc posted on Canvas before the end of the week

Workshop Days (Thurs)

Use your own computers. Have the latest R version installed.

This will take LONGER than the in-class time. Start the workshops prior to class. Use class time to discuss with your peers and instructor.

You must submit exercises (9 out of 10) each Monday by 11:59 pm. No **need** to submit week #1 ☺

Three minor assignments (40%)

1. Explore, explain, and summarize your dataset
   - Due week 3
   - 10%
2. Analyze a linear model
   - Due week 6
   - 15%
3. Compare and extend your models
   - Due week 12
   - 15%

Late policy: 10% per day

Final Project (20%)

- This culminating assignment challenges you to conduct original research using your dataset and present your findings in the format of a scientific paper.
- You are expected to apply statistical tools covered throughout the course and clearly interpret your results in the context of your research questions.

Late policy: 10% per day

Attendance and participation (10%)

- 2 excused absences
- 3-5 absences lose 5%
- More than 5 absences lose 10%

Where to find data to learn R and do assignments?

- This class will be most useful for you if you use your OWN DATA
- Ask your lab members for their datasets that are published
- Many online journals require database submission (QA/QC is variable)
- Manually enter data from published papers figs and tables
- No copyright on published data (it's ok to use for R practice or meta-analysis)
- Graphics tool: https://www.datathief.org/
- Online data archives, e.g., Ecological Archives, https://esapubs.org/archive/, Genbank, Dryad (https://datadryad.org/search)
- Permissions/conditions may be required to publish results from archives.

Creating an inclusive and respectful classroom is important

- **All** voices are important in lecture and discussions
- We are all at different points in learning R and have different strengths
- We will work together to create a safe inclusive learning space by brainstorming guidelines for communication, participation, and discussions
- Examples: raise hand to talk, don’t interrupt

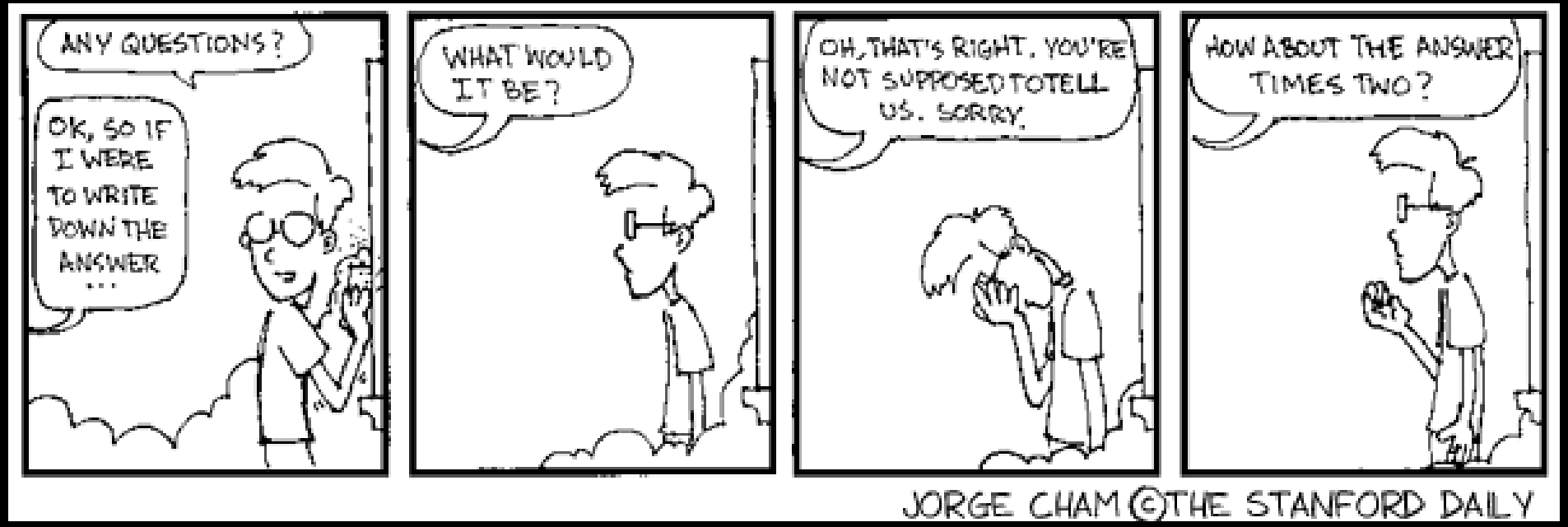
Discussion

What are some specific group guidelines for us to create a respectful, inclusive space in lecture and discussions where ***all*** voices have input?

Email and Office Hours

Office Hours:
Monday 10:30-11:30 and Thursday 4:00-5:00 in SHB 541. Also available via zoom.

Please use email to contact me outside of office hours. I will typically respond within 48 hours of receiving your message.



Well-Being and Health Are Important

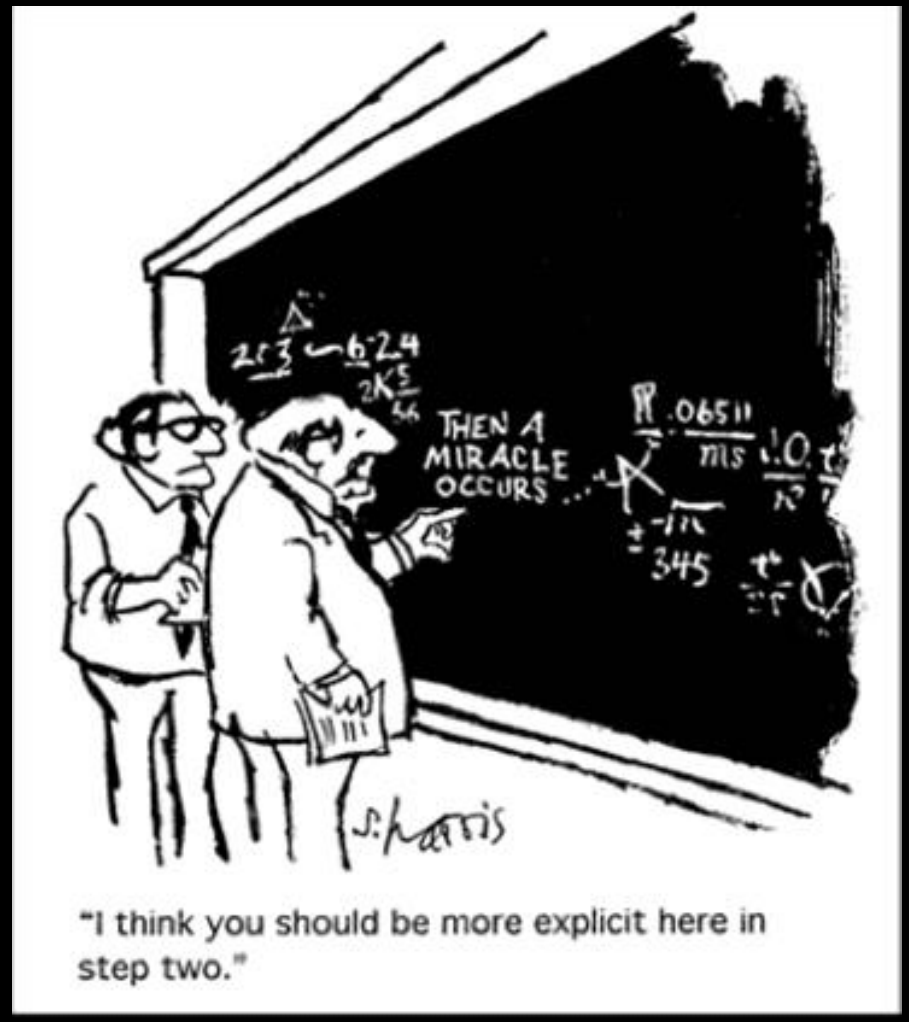
- Don’t come to class if you are sick
- Good-posture/desk setup while coding is important
- Work-life balance is important. Enjoy what makes you happy.
- Come talk to me if you have any concerns

See Syllabus for additional resources

Teaching Philosophy

Active learning requires your effort.

- I won't do it for you.
- Using code you don't understand is ultimately useless



"I think you should be more explicit here in step two."

Less guidance is given in a grad course than an undergrad course.

Syllabus Policy Statements

- ACADEMIC INTEGRITY
- USE OF GENERATIVE ARTIFICIAL INTELLIGENCE (AI)
- COPYRIGHT INFRINGEMENT
- COURSE TIME COMMITMENT
- NONDISCRIMINATION AND ANTI-HARASSMENT
- TITLE IX
- ACCESSIBILITY
- RESPONSIBLE CONDUCT OF RESEARCH
- MISCONDUCT IN RESEARCH

Discussion

What is “acceptable” use of generative AI in this class and in your research?

Outline for today

- Meet your Instructor
- About the course
- Why we use R
- Organizing data for analysis in R
- Wrap-up
  - To Do List before Thursday's workshop
  - Pre-course survey
  - Sign up for Discussion (2 students per presentations – on canvas)

What is R?

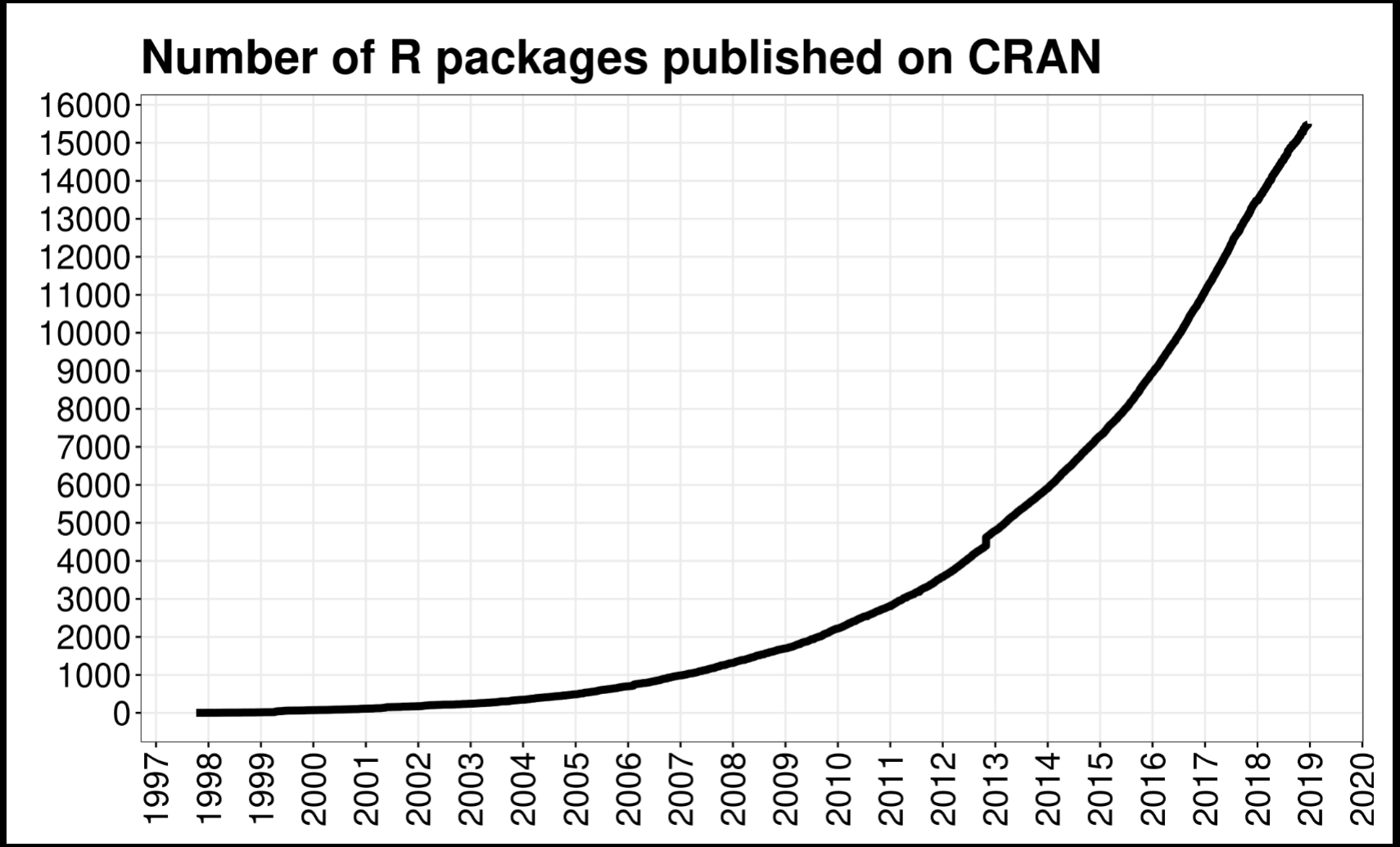
- R is a language and environment for statistical computing and graphics.
- Free, collaborative
- The current R is the result of a collaborative effort involving contributors from all over the world.

Discussion

Brainstorm “Good” and “Bad” things about using R

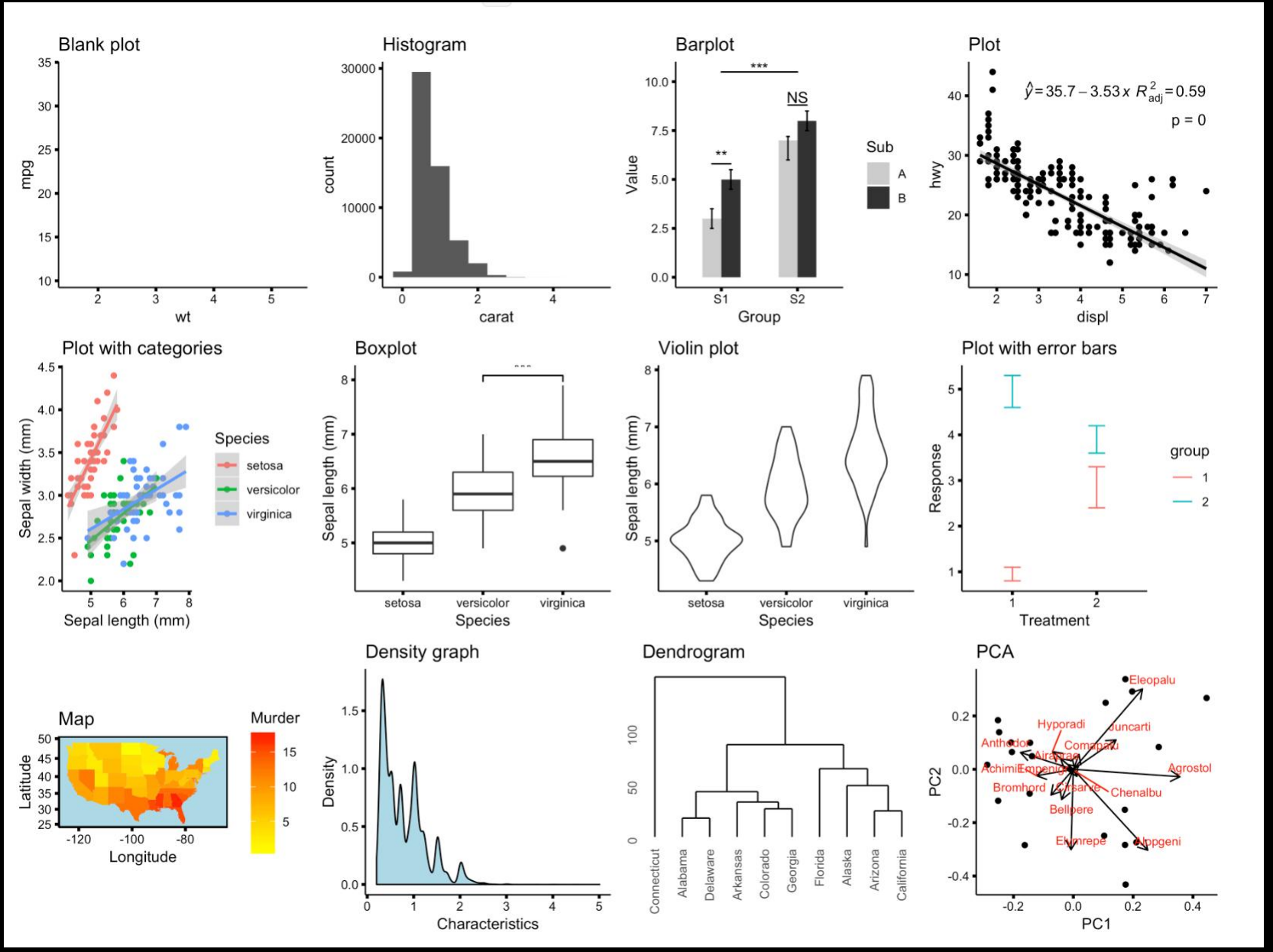
Why use R?

- Free
- More scientists use it every year!
- More than 16000 packages registered within CRAN



Why use R?

A lot features: customizable graphs



Outline for today

- Meet your Instructor
- About the course
- Why we use R
- Organizing data for analysis in R
- Wrap-up
  - To Do List before Thursday's workshop
  - Pre-course survey
  - Sign up for Discussion (2 students per presentations – on canvas)

How will you organize and standardize R Scripts and Data?

Think about **long-term organization and storage** of R scripts, data, outputs, figures

- Your computer may change during grad school and afterwards
- I have had 8 different laptops since my MSc and can still find and run R code from 2008 (although sometimes it is difficult)
- **How?** organization of files and **lots of annotations in scripts**
- It can be years between data analysis and journal revisions (redo analysis, edit figures)

How will you organize and standardize R Scripts and Data?

Think about **long-term organization and storage** of R scripts, data, outputs, figures

- For your thesis, I'd recommend organizing by data chapters which is equivalent to published papers.
  - 1 pre-processing script
  - 1 Main analysis script with full figures in it (stats, models on full dataset)
  - 1 separate script for only final figures (at the end)
  - Archiving data and analysis is **often** required

Preparing data for R

- Datasets should be stored as comma separated files (.csv)
- Naming data files:
  - Use descriptive informative names (final.csv = bad)
  - Avoid numbers
  - Don’t separate names with dots
- Naming variables:
  - Use short informative variables (“time_1” not “first time measurement”)
  - Case sensitive (use lowercase)
- Common mistakes:
  - Text in numeric columns, spaces, typos, use NA or blanks (not -9999)

Best practices for data entry

What to enter in columns:

- Use brief, informative variable names in plain text. Keep more detailed explanations of variables in a separate text file.
- Avoid spaces in variable names – use a dot or underscore instead (e.g., size.mm or size_mm).
- Leave missing data cells blank.
- Avoid non-numeric characters in columns of numeric data. R will assume that the entire column is non-numeric. For example, avoid using a question mark “12.67?” to indicate a number you are not sure about. Put the question mark and other comments into a separate column just for comments.
- Use the international date format (YYYY-MM-DD).
- Keep commas out, because they are column delimiters in your .csv file
- R is case sensitive
- A “long” layout is recommended, instead of a “wide” layout, when using linear models to analyze data. Use different columns for variables and different rows for sample units.

Bring the.csv of your thesis dataset to a workshop, I’m happy to look at the headers, and layout of it ***before*** you import it into R

How would you assess this database? Is it ”R friendly”?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Animal	Day	Month	Year	Depth	Body Mass	Rep	Single Dive	Dive Bout	Dive.Duration	tmrs	tmrs.bin	total.time	Heart Rate Beats/Min	Average Metabolic Rate L O2/Min
2	Boni	7	aug	2008	10	142	1	Single		3.95	4.39	4 plus	8.85	53.24	1.801132156
3	Boni	14	aug	2008	10	144.5	2	Single		4.85	3.41	2 to 4	8.91	52.15	1.664475516
4	Boni	16	sep	2008	40	142	3	Single		7.44	2.26	2 to 4	10.72	85.47	2.037451206
5	Boni	19	aug	2008	40	155.5	1	Single		6.87	2.01	2 to 4	6.87	47.34	1.587817682
6	Hazy	5	aug	2008	10	171.5	1	Single		2.64	4.48	4 plus	7.84	97.23	2.091133827
7	Hazy	7	aug	2008	10	172	2	Single		3.69	2.25	2 to 4	6.35	89.81	2.328101564
8	Hazy	14	aug	2008	10	173	3	Single	Dive Bout	4.39	2.85	2 to 4	7.68	78.64	2.055794036
9	Hazy	4	sep	2008	40	171.5	3		Dive Bout	5.91	3.35	2 to 4	10.34	68.78	1.960670476
10	Sitka	1	aug	2008	10	223.5	1		Dive Bout	4.73	5.29	4 plus	10.44	58.07	1.792121579
11	Sitka	7	aug	2008	10	223	2		Dive Bout	4.2	0.45	less than 1	10.44	53.06	1.727073004
12	Sitka	4	sep	2008	40	215	1		Dive Bout	6.48	5.33	4 plus	12.97	45.92	2.268856079
13															

Outline for today

- Meet your Instructor
- About the course
- Why we use R
- Organizing data for analysis in R
- Wrap-up
  - To Do List before Thursday's workshop
  - Sign up for Discussion (2 students per presentations – on canvas)
  - Pre-course survey

To do before workshop this Thurs

- Have latest R version installed (Rstudio preferred)
- Set your R aesthetics and window panes as you prefer
  - Rstudio/Preferences/Appearance
  - Rstudio/Preferences/Pane Layout
- Set up ”your system” of how you will organize R scripts, figs, tables etc

- Workshop 1: Intro to R
- Two Bonus workshops:
  - Data visualization using ggplot
  - Power analysis

First discussion paper

- Zuur: A protocol for data exploration to avoid common statistical problems.
- Need two presenters for next week:
  - Sign up in Canvas!

Pre-course survey posted on Canvas

- Who are you academically?
- Who are you outside of your thesis?
- Familiarity with R?
- Familiarity with stats theory?
- What are some potential types of data you want to collect for your thesis?
- Types of analysis you are interested in learning in R?

Anonymous Feedback via Canvas – anytime!