ES 193DS homework 3

Katelyn Vidal

2024-06-02

### Set up:

#reading in packages  
# general use  
library(tidyverse)  
library(readxl)  
library(here)  
library(janitor)  
  
# visualizing pairs  
library(GGally)  
  
# model selection  
library(MuMIn)  
  
# model predictions  
library(ggeffects)  
  
# model tables  
library(gtsummary)  
library(flextable)  
library(modelsummary)  
library(ModelMetrics)  
  
# read in data  
drought\_exp <- read\_xlsx(path = here("data",   
 "Valliere\_etal\_EcoApps\_Data.xlsx"),  
 sheet = "First Harvest")  
  
# quick look at data   
str(drought\_exp)

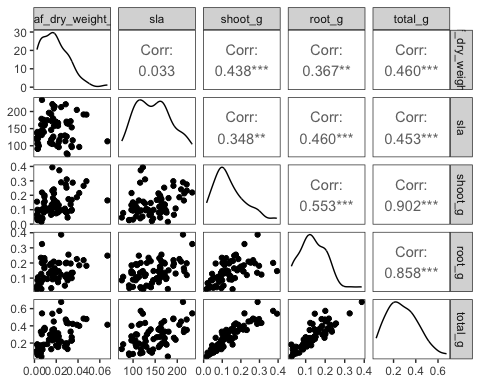
tibble [70 × 13] (S3: tbl\_df/tbl/data.frame)  
 $ Species : chr [1:70] "ENCCAL" "ENCCAL" "ENCCAL" "ENCCAL" ...  
 $ Water : chr [1:70] "WW" "WW" "WW" "WW" ...  
 $ Rep # : num [1:70] 1 2 3 4 5 1 2 3 4 5 ...  
 $ Height (cm) : num [1:70] 5.8 4.9 8.4 6.5 7.1 3.2 4.4 4.2 4.5 3.9 ...  
 $ Leaf # : num [1:70] 11 8 11 12 10 7 7 10 8 6 ...  
 $ Leaf dry weight (g): num [1:70] 0.0294 0.0185 0.0177 0.0178 0.0164 0.017 0.0193 0.0153 0.0159 0.0133 ...  
 $ Leaf area (cm2) : num [1:70] 5.01 3.98 3.69 3.84 3.63 3.06 3.1 2.94 2.73 2.61 ...  
 $ SLA : num [1:70] 170 215 209 216 222 ...  
 $ Total LA : num [1:70] 55.1 31.8 40.6 46.1 36.3 ...  
 $ Shoot (g) : num [1:70] 0.253 0.164 0.241 0.213 0.232 ...  
 $ Root (g) : num [1:70] 0.202 0.165 0.209 0.146 0.12 ...  
 $ Total (g) : num [1:70] 0.455 0.329 0.45 0.359 0.352 ...  
 $ R:S : num [1:70] 0.8 1 0.9 0.7 0.5 0.8 1.2 3.1 0.9 1.2 ...

class(drought\_exp)

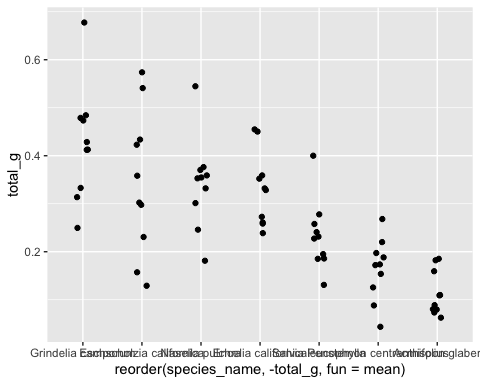
[1] "tbl\_df" "tbl" "data.frame"

# cleaning  
drought\_exp\_clean <- drought\_exp %>%   
 # nicer column names  
 clean\_names() %>%   
 # adding column with species scientific names  
 mutate(species\_name = case\_when(   
 species == "ENCCAL" ~ "Encelia californica", # bush sunflower  
 species == "ESCCAL" ~ "Eschscholzia californica", # California poppy  
 species == "PENCEN" ~ "Penstemon centranthifolius", # Scarlet bugler  
 species == "GRICAM" ~ "Grindelia camporum", # great valley gumweed  
 species == "SALLEU" ~ "Salvia leucophylla", # Purple sage  
 species == "STIPUL" ~ "Nasella pulchra", # Purple needlegrass  
 species == "LOTSCO" ~ "Acmispon glaber" # deerweed  
 )) %>%   
 # moving species\_name column after species  
 relocate(species\_name, .after = species) %>%   
 # adding column with full treatment names  
 mutate(water\_treatment = case\_when(   
 water == "WW" ~ "Well watered",  
 water == "DS" ~ "Drought stressed"  
 )) %>%   
 # moving water\_treatment column after water  
 relocate(water\_treatment, .after = water)

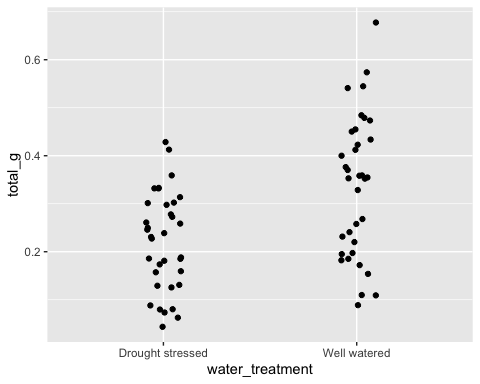
# visualizing correlations with ggpairs  
#| fig-height: 8  
#| fig-width: 8  
  
# data frame  
ggpairs(drought\_exp\_clean,   
 # columns to visualize  
 columns = c("leaf\_dry\_weight\_g",   
 "sla",   
 "shoot\_g",   
 "root\_g",   
 "total\_g"),   
 # calculating Pearson correlation coefficient  
 upper = list(method = "pearson")) +   
 # cleaner theme  
 theme\_bw() +   
 # getting rid of gridlines  
 theme(panel.grid = element\_blank())



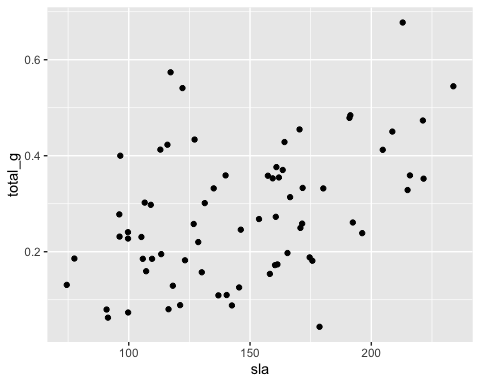
# visualizing mass as a function of species  
#| fig-width: 10  
#| fig-height: 5  
  
# data frame  
ggplot(data = drought\_exp\_clean,   
 # reordering x-axis  
 aes(x = reorder(species\_name,  
 # in reverse order of mean total mass  
 -total\_g,  
 # calculating mean to reorder  
 fun = mean),  
 # y-axis  
 y = total\_g)) +  
 # narrow jitter  
 geom\_jitter(width = 0.1,  
 # not jittering points up and down  
 height = 0)



# visualizing mass a a function of water treatment  
# data frame  
ggplot(data = drought\_exp\_clean,   
 # x-axis  
 aes(x = water\_treatment,  
 # y-axis  
 y = total\_g)) +   
 # narrow jitter  
 geom\_jitter(width = 0.1,   
 # not jittering points up and down  
 height = 0)

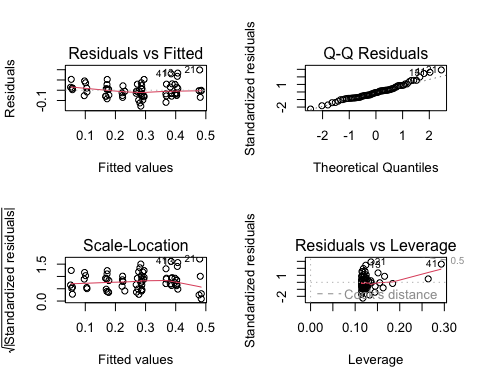


# visualizing mass as a function of sla  
# data frame  
ggplot(data = drought\_exp\_clean,  
 # x-axis  
 aes(x = sla,   
 # y-axis  
 y = total\_g)) +   
 # scatterplot  
 geom\_point()

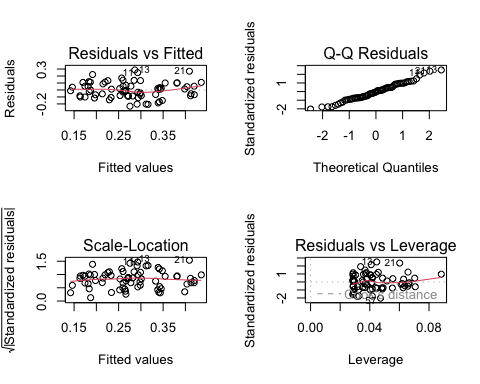


# null model  
model0 <- lm(total\_g ~ 1,   
 data = drought\_exp\_clean)

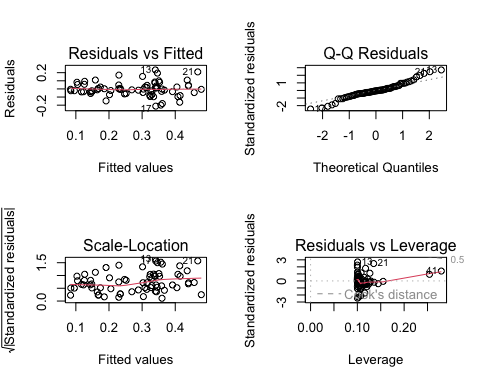
# model 1, saturated model, total biomass as a function of SLA, water treatment, and species  
model1 <- lm(total\_g ~ sla + water\_treatment + species\_name,  
 data = drought\_exp\_clean)  
# show plots 2x2  
par(mfrow = c(2, 2))  
# show model 1  
plot(model1)



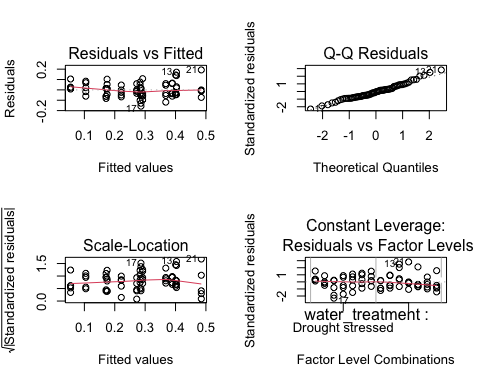
# model 2, total biomass as a function of SLA and water treatment  
model2 <- lm(total\_g ~ sla + water\_treatment,  
 data = drought\_exp\_clean)  
  
par(mfrow = c(2, 2))  
# show model 2 diagnostics  
plot(model2)



# model 3, total biomass as a function of SLA and species  
model3 <- lm(total\_g ~ sla + species\_name,  
 data = drought\_exp\_clean)  
  
par(mfrow = c(2, 2))  
# show model 3 diagnostics  
plot(model3)



# model 4, total biomass as a function of water treatment and species  
model4 <- lm(total\_g ~ water\_treatment + species\_name,  
 data = drought\_exp\_clean)  
  
par(mfrow = c(2, 2))  
# show model 4 diagnostics  
plot(model4)



# model selection  
model.sel(model0,  
 model1,   
 model2,   
 model3,  
 model4)

Model selection table   
 (Int) sla spc\_nam wtr\_trt df logLik AICc delta weight  
model4 0.05455 + + 9 88.598 -156.2 0.00 0.772  
model1 0.07994 -0.0002475 + + 10 88.741 -153.8 2.44 0.228  
model3 -0.03315 0.0012900 + 9 72.538 -124.1 32.12 0.000  
model2 0.04670 0.0012810 + 4 52.220 -95.8 60.37 0.000  
model0 0.27900 2 39.580 -75.0 81.22 0.000  
Models ranked by AICc(x)

# summary of model 4  
summary(model4)

Call:  
lm(formula = total\_g ~ water\_treatment + species\_name, data = drought\_exp\_clean)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.157087 -0.046953 -0.003733 0.041244 0.192657   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.05455 0.02451 2.225 0.02973 \*   
water\_treatmentWell watered 0.11695 0.01733 6.746 5.90e-09 \*\*\*  
species\_nameEncelia californica 0.21774 0.03243 6.714 6.70e-09 \*\*\*  
species\_nameEschscholzia californica 0.23164 0.03243 7.143 1.22e-09 \*\*\*  
species\_nameGrindelia camporum 0.31335 0.03243 9.662 5.53e-14 \*\*\*  
species\_nameNasella pulchra 0.22881 0.03243 7.055 1.72e-09 \*\*\*  
species\_namePenstemon centranthifolius 0.05003 0.03243 1.543 0.12799   
species\_nameSalvia leucophylla 0.12020 0.03243 3.706 0.00045 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.07252 on 62 degrees of freedom  
Multiple R-squared: 0.7535, Adjusted R-squared: 0.7257   
F-statistic: 27.08 on 7 and 62 DF, p-value: < 2.2e-16

### Problem 1. Multiple linear regression: model selection and construction (52 points)

Use the information from the homework-starter-doc.qmd to do this problem.

#### a. Make a table *or* list of all the models from class and the last one you constructed on your own. Write a caption for your table. (8 points)

**Caption:** table captions typically go above the table. Number the table and provide a title. Describe what is in the table (columns and rows).

**Table:** In your table, each row should be a model with the model number (1, 2, 3, etc.) and the predictors for each model.

# making a matrix called 'all\_models'  
all\_models <- tribble(  
 # specify column titles  
 ~model, ~intercept, ~sla, ~well\_watered, ~Encelia\_californica, ~Eschscholzia\_californica, ~Grindelia\_camporum, ~Nasella\_pulchra, ~Penstemon\_centranthifolius, ~Salvia\_leucophylla, ~AIC, ~deltaAIC,  
 # specify row names and values  
 "null", 0.279, NA, NA, NA, NA, NA, NA, NA, NA, -75.2,81.22,  
 "model 1", 0.080, 0, 0.122, 0.238, 0.234, 0.330, 0.241, 0.061, 0.117, -157.5, 2.44,  
 "model 2", 0.047, 0.001, 0.090, NA, NA, NA, NA, NA, NA, -96.4, 60.37,  
 "model 3", -0.033 , 0.001, NA, 0.115, 0.222, 0.226, 0.168, -0.006, 0.139, -127.1, 32.12,  
 "model 4", 0.055, NA, 0.0117, 0.218, 0.232, 0.313, 0.229, 0.050, 0.120, -159.2, 0  
)  
  
# display matrix   
all\_models

# A tibble: 5 × 12  
 model intercept sla well\_watered Encelia\_californica Eschscholzia\_califor…¹  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 null 0.279 NA NA NA NA   
2 model… 0.08 0 0.122 0.238 0.234  
3 model… 0.047 1e-3 0.09 NA NA   
4 model… -0.033 1e-3 NA 0.115 0.222  
5 model… 0.055 NA 0.0117 0.218 0.232  
# ℹ abbreviated name: ¹​Eschscholzia\_californica  
# ℹ 6 more variables: Grindelia\_camporum <dbl>, Nasella\_pulchra <dbl>,  
# Penstemon\_centranthifolius <dbl>, Salvia\_leucophylla <dbl>, AIC <dbl>,  
# deltaAIC <dbl>

# turning everything into a flextable  
 flextable(all\_models) %>%   
 # changing the column names to look nicer  
 set\_header\_labels(well\_watered = "Well watered",  
 Encelia\_californica = " Encelia californica",  
 Eschscholzia\_californica = "Eschscholzia californica",  
 Grindelia\_camporum = "Grindelia camporum,",  
 Nasella\_pulchra = "Nasella pulchra",  
 Penstemon\_centranthifolius = "Penstemon centranthifolius",  
 Salvia\_leucophylla = "Salvia leucophylla",  
 deltaAIC = "delta AIC") %>%   
 # adding headers  
 add\_header\_row(colwidths = c(2, 1, 1, 3, 3, 2),  
 values = c("", "", "Water treatment","Species", "", "")) %>%  
 # adding caption  
 set\_caption(caption = as\_paragraph(colorize("Table 1. Rows are models and columns show the intercept, predictors (sla, water treatment, and species), AIC, and ∆AIC. Cells for each predictor contain model estimate values.", color = "black"))) %>%   
 # changing theme and width of table  
 theme\_vanilla() %>%   
 width(width = 10)

|  | |  | **Water treatment** | **Species** | | |  | | |  | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **model** | **intercept** | **sla** | **Well watered** | **Encelia californica** | **Eschscholzia californica** | **Grindelia camporum,** | **Nasella pulchra** | **Penstemon centranthifolius** | **Salvia leucophylla** | **AIC** | **delta AIC** |
| null | 0.279 |  |  |  |  |  |  |  |  | -75.2 | 81.22 |
| model 1 | 0.080 | 0.000 | 0.1220 | 0.238 | 0.234 | 0.330 | 0.241 | 0.061 | 0.117 | -157.5 | 2.44 |
| model 2 | 0.047 | 0.001 | 0.0900 |  |  |  |  |  |  | -96.4 | 60.37 |
| model 3 | -0.033 | 0.001 |  | 0.115 | 0.222 | 0.226 | 0.168 | -0.006 | 0.139 | -127.1 | 32.12 |
| model 4 | 0.055 |  | 0.0117 | 0.218 | 0.232 | 0.313 | 0.229 | 0.050 | 0.120 | -159.2 | 0.00 |

#### b. Write a 5-6 sentence “statistical methods” section. (8 points)

Your answer should be in paragraph form and include:

* how you addressed the central question(s) (i.e. to examine the influence of \_\_\_\_, \_\_\_\_, and \_\_\_\_ on \_\_\_\_\_, I…)
* how you chose the final model (i.e. to determine the model that best described \_\_\_\_, I…)
* how you visually determined that your final model conformed to the assumptions of a linear model (i.e. to evaluate linear model assumptions, I…)

**To examine the influence of specific leaf area, water treatment, and species on total biomass, I used a linear model to test my null hypothesis, which was that these variables do not predict total biomass, and my saturated model, which was that all three variables predict total biomass. To determine the model that best described total biomass, I used the Multi-Model Inference (MuMIn) package and looked at which model contained the lowest AIC, which was model 4. To evaluate linear model assumptions, I looked at the diagnostic plots to visually assess for normality and homoscedasticity and used Cook’s distance to determine that no outliers were influencing my model predictions.**

#### c. Make a visualization of the model predictions with underlying data for your “best” model. (20 points)

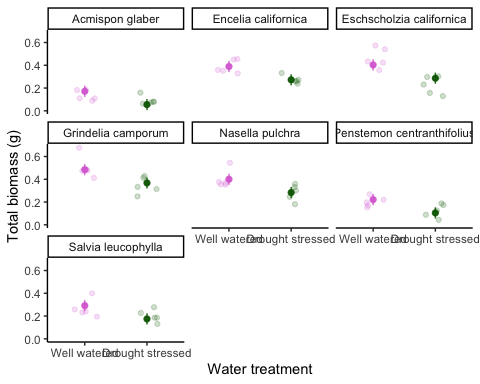
Show and annotate all your code. For full credit:

* make the underlying data more transparent than the model predictions
* display species names in full (not as species codes like ENCCAL or ESCCAL)
* display water treatment types in full (not as WW or DS)
* represent well-watered and drought stressed treatments with different colors
* use colors that are *not* the default ggplot() colors
* facet your plot by species
* remove the legend
* finalize the plot

# model predictions for model 4  
model\_preds <- ggpredict(model4,   
 terms = c("water\_treatment",   
 "species\_name"))

# creating new data frame of model predictions for plotting  
model\_preds\_for\_plotting <- model\_preds %>%   
 # renaming columns to make this easier to use  
 rename(water\_treatment = x,  
 species\_name = group)

# visualization of the model predictions with underlying data for model 4  
ggplot() +  
 # underlying data  
 geom\_jitter(data = drought\_exp\_clean,  
 # x-axis  
 aes(x = water\_treatment,  
 #y-axis  
 y = total\_g,  
 # color by water\_treatment  
 color = water\_treatment),  
 # make points more transparent  
 alpha = 0.2,  
 # make jittered points narrower  
 width = 0.2) +  
 # model predictions with 95% CI   
 geom\_pointrange(data = model\_preds\_for\_plotting,  
 aes(x = water\_treatment,   
 y = predicted,  
 ymin = conf.low,  
 ymax = conf.high,  
 color = water\_treatment),  
 alpha = 0.9,  
 size = 0.3) +  
 # change color of model predictions with 95% CI and underlying data  
 scale\_color\_manual(values = c("Well watered" = "orchid", "Drought stressed" = "darkgreen")) +  
 # cleaner theme  
 theme\_classic() +  
 # creating different panels for species  
 facet\_wrap(~species\_name) +  
 # rename x and y axes  
 labs(x = "Water treatment",  
 y = "Total biomass (g)") +  
 # remove legend  
 theme(legend.position = "none")



#### d. Write a caption for your visualization. (6 points)

**Figure 1. Predicted values of total biomass as a function of water treatment and species**  
**Data from ‘Dryad’ package (Valliere J, Zhang J, Sharifi M., Rundel P, 2019). Jittered points represent observations of the total biomass (g) of species in well watered and drought stressed water treatments, (total n = 70). The center dot represents the model prediction of total biomass of species for each water treatment, and the whiskers represent 95% CI. Colors represent water treatment type (pink = well watered, green = drought stressed).**

#### e. Write a 3-4 sentence results section. (10 points)

Your answer should be in paragraph form and address the following points:

* what predictors “best” described total mass (include model statistics here)?
* on average, what differences did you find between water treatments?
* on average, what differences did you find between species?

**Results:**  
**We found that water treatments and species significantly predicted total biomass, but not SLA (Table 1). Across species on average, plants in well watered treatment are 0.12 ± 0.017 g greater than plants in drought stressed treatment. Across water treatments on average, Eschscholzia californica plants are 0.23 ± 0.032 g more than Acmispon glaber (F(7, 62) = 27.08, p < 0.001, ⍺ = 0.05, adjusted R2 = 0.73).**

### Problem 2. Affective visualization (24 points)

In this problem, you will create an **affective visualization** using your personal data in preparation for workshop during week 10.

In lecture, we talked about the three vertices of data visualization: 1) exploratory, 2) affective, and 3) communicative. We’ve done a lot of exploratory and communicative visualization, but have yet to think about affective visualization.

When thinking of affective visualization, you can expand your ideas of what data visualization *could* be. Some examples of affective visualizations include:

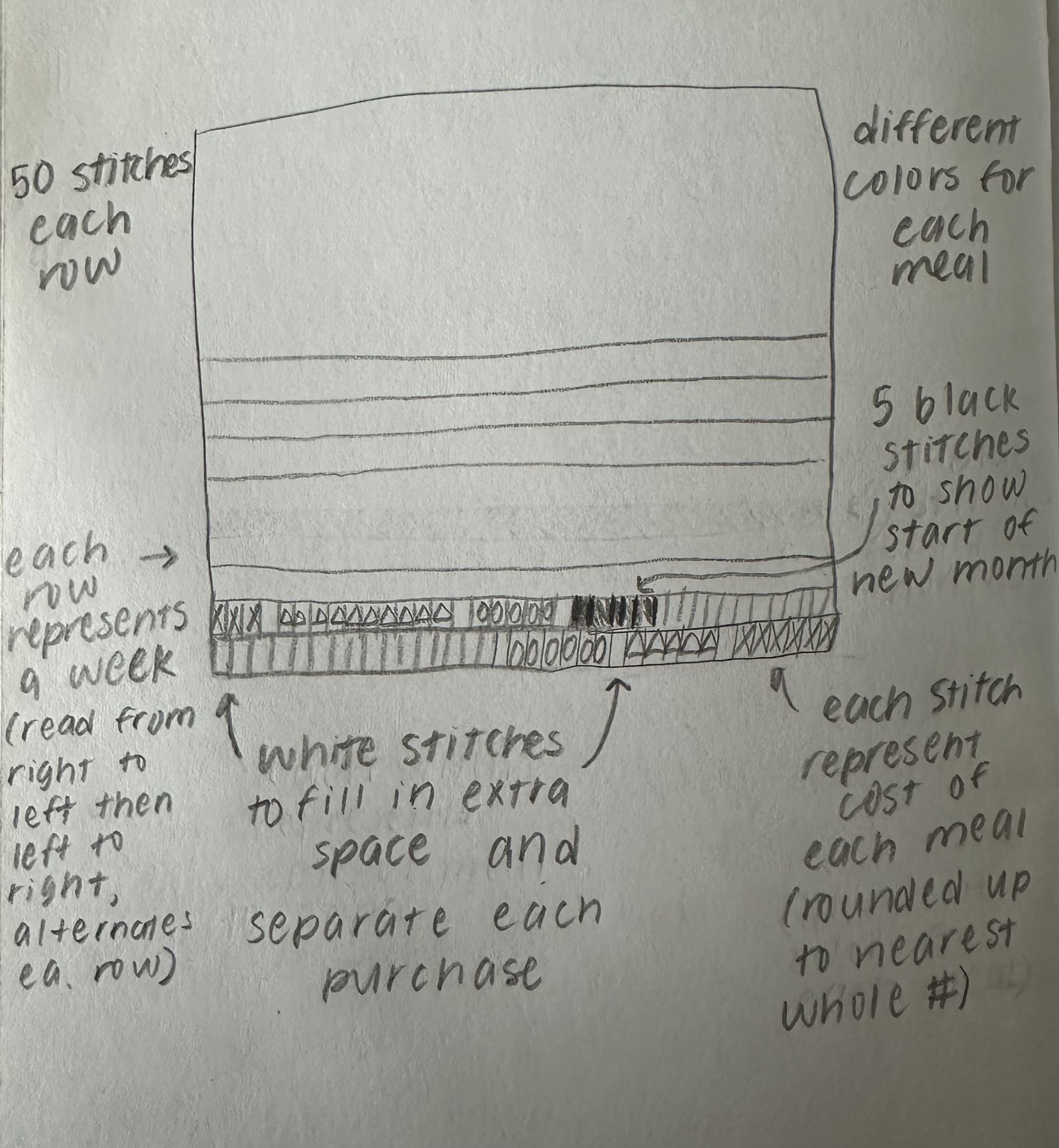
* [Jill Pelto’s paintings](https://www.jillpelto.com/gallery-1)
* [Lorraine Woodruff-Long’s warming strips quilt](https://www.instagram.com/p/C689Kybxget/?utm_source=ig_web_copy_link&igsh=MzRlODBiNWFlZA==)
* [Stefanie Posavec and Giorgia Lupi’s Dear Data project](https://www.dear-data.com/all)

**Before starting, update your spreadsheet of observations.**

#### a. Describe in words what an affective visualization could look like for your personal data (3-5 sentences). (2 points)

**For my personal data, I could crochet a square, with each stitch representing how much I spend for each meal, and use different colors to represent the meal type. I could also have each row represent each week, and use white yarn to fill in the remaining stitches and separate each purchase (each row will have 50 stitches). This will allow me to see which meals I tend to spend the most money on and see which weeks I spent the most and least amount of money. I could also separate each month by using a contrasting yarn color.**

#### b. Create a sketch (on paper) of your idea. (2 points)



#### c. Make a draft of your visualization. (12 points)

Feel free to be creative with this! *You do not have to do this in R.* You could create a sculpture, painting, textile object, etc.

If you are making your visualization in R, show the annotated code and the output.

If you are making your visualization outside of R, include a photo of your visualization in your submission.



#### d. Write an artist statement. (8 points)

An artist statement gives the audience context to understand your work. Write 4-5 sentences to address:

* the content of your piece (what are you showing?)
* the influences (what did techniques/artists/etc. did you find influential in creating your work?)
* the form of your work (written code, watercolor, oil painting, etc.)
* your process (how did you create your work?)

**This crochet piece represents my weekly and monthly spending habits on food, with each stitch representing the cost of a meal (USD, rounded up to the nearest whole number) and each row representing each week, with a series of 5 black stitches to represent the start of a new month. The colors represent the type of meal (blue = lunch, light brown = snack, dark brown = dinner), and the white yarn is used to fill the remaining space in each row (rows contain 50 stitches each) and separate each purchase. This piece is influenced by Lorraine Woodruff-Long’s warming strips quilt representing global temperature change, which inspired me to show how my spending habits on food changed over time. I created this piece by first calculating what my max cost of food was each week to determine how many stitches to include in each row. I then started crocheting, adding in each color, filler, and divider until I reached the end of my observations.**

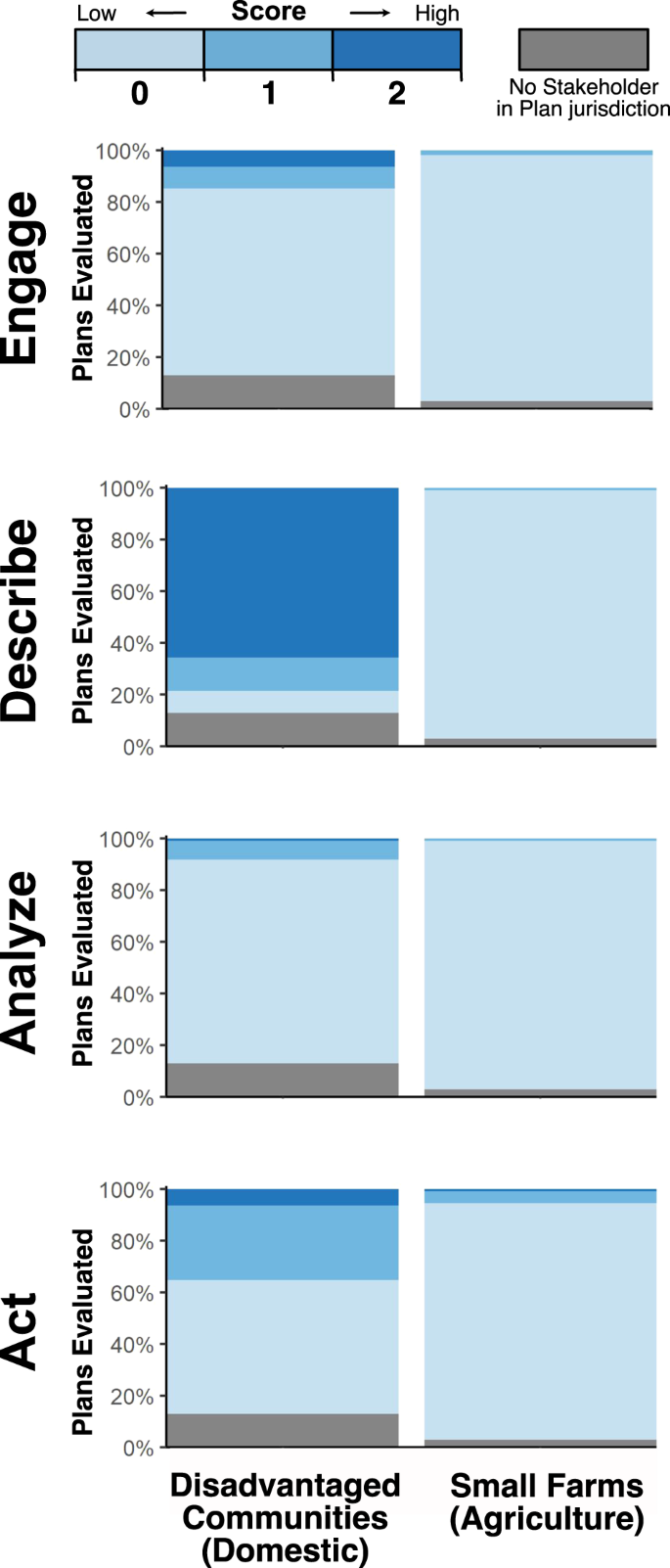
### Problem 3. Statistical critique (36 points)

At this point, you have seen and created a lot of figures for this class. Revisit the paper you chose for your critique and your homework 2, where you described figures or tables in the text. Address the following in full sentences (3-4 sentences each).

**For this section of your homework, you will be evaluated on the logic, conciseness, and nuance of your critique.**

#### a. Revisit and summarize (6 points)

What are the statistical tests the authors are using to address their main research question?

**The authors used Mann–Whitney Wilcoxon Rank-Sum tests to address their main research question: “How well stakeholders are included in and protected from groundwater depletion in Sustainability Plans under California’s Sustainable Groundwater Management?”**  

#### b. Visual clarity (10 points)

How clearly did the authors *visually* represent their statistics in figures? For example, are the x- and y-axes in a logical position? Do they show summary statistics (means and SE, for example) and/or model predictions, and if so, do they show the underlying data?

**The authors clearly labeled their x- and y-axes and labeled each subplot in a larger font to distinguish that the plots show different integration components. They also use contrasting colors to represent each integration score that also follows a blue theme, making the statistics visually pleasing and easy to read. They also provide a legend that shows readers what each color represents and what color/score is associated with being high, moderate, and low.**

#### c. Aesthetic clarity (10 points)

How well did the authors handle “visual clutter”? How would you describe the the data:ink ratio?

**There is minimal “visual clutter” in this figure as there are no grid lines or redundant information that make the plot difficult to read. The bars also clearly represent each score using colors that are easy to distinguish between. However, the title of the subplots seems out of place, being next to the y-axis title. The legend is also clearly labeled and in a logical location that doesn’t interfere with the data in the plots. I would describe the data:ink ratio as fairly high because the labels are informative, the percentages are lightened while the titles are in bold, which creates a contrast that helps reduce clutter, there’s no clutter in the background, and the bars are clearly separated.**

#### d. Recommendations (can be longer than 4 sentences, 10 points)

What recommendations would you make to make the figure better? What would you take out, add, or change? Provide explanations/justifications for each of your recommendations.

**Alternatively**, if they did not represent their statistics in a figure, what kind of figure would you recommend to them? Describe the x- and y-axes, with any geometries and aesthetics (colors, etc.). Provide enough explanation such that someone would be able to create the figure you describe in code. Provide explanations/justifications for the figure you would recommend.

**To make the figure better, I would label each subdivision of the bars with the percentage it represents. It can be misleading to look at the top of each subdivision to determine the percentage because they don’t all start at 0%. So, labeling each subdivision allows people to look at the data without having to do calculations to determine the percentage. I would also put the title of the subplots on top of each subplot so people don’t confuse the title with the y-axis title or get confused about which title represents the y-axis. I would also organize the plots in a 2x2 format rather than a 1x4 format so that you can easily compare the plots without having to scroll down to look at all the plots together.**