Homework-03

Ada Chibueze

2024-05-30

#Link to Github repo: https://github.com/adachibueze/ENVS193DS_homework-08

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)

#reading in the data
drought_exp <- read_xlsx(path = here("data",
                                     "Valliere_etal_EcoApps_Data.xlsx"),
                         sheet = "First Harvest")
```

cleaning

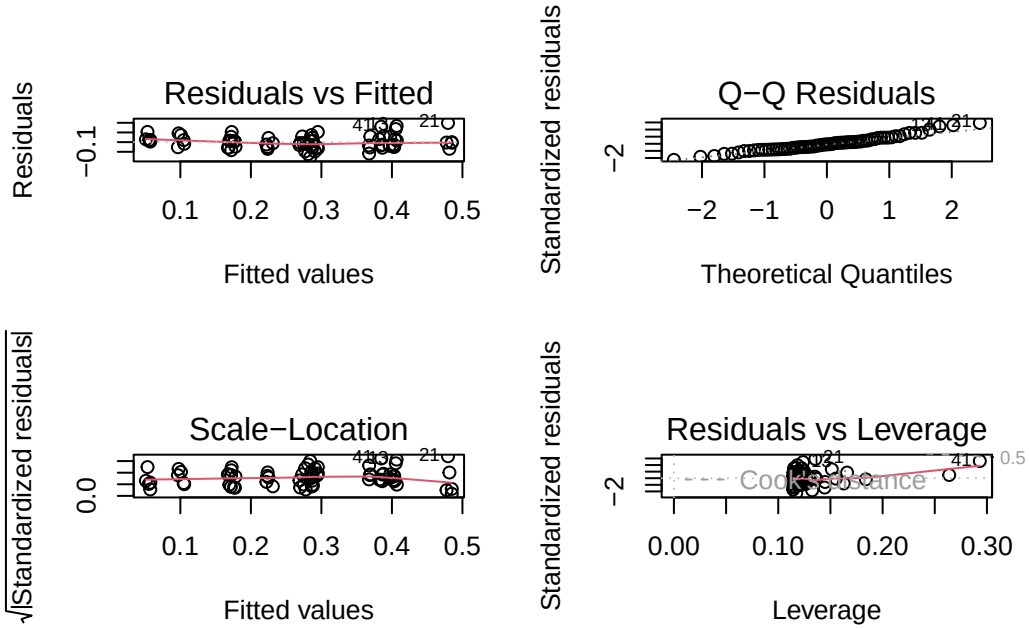
```
# cleaning
drought_exp_clean <- drought_exp %>%
  clean_names() %>% # nicer column names
  mutate(species_name = case_when( # adding column with species scientific names
    species == "ENCCAL" ~ "Encelia californica", # bush sunflower
    species == "ESCCAL" ~ "Eschscholzia californica", # California 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
  )) %>%
  relocate(species_name, .after = species) %>% # moving species_name column after species
  mutate(water_treatment = case_when( # adding column with full treatment names
    water == "WW" ~ "Well watered",
    water == "DS" ~ "Drought stressed"
  )) %>%
  relocate(water_treatment, .after = water) # moving water_treatment column after water
```

0. Null model

```
model0 <- lm(total_g ~ 1, # formula
             data = drought_exp_clean) # data frame
```

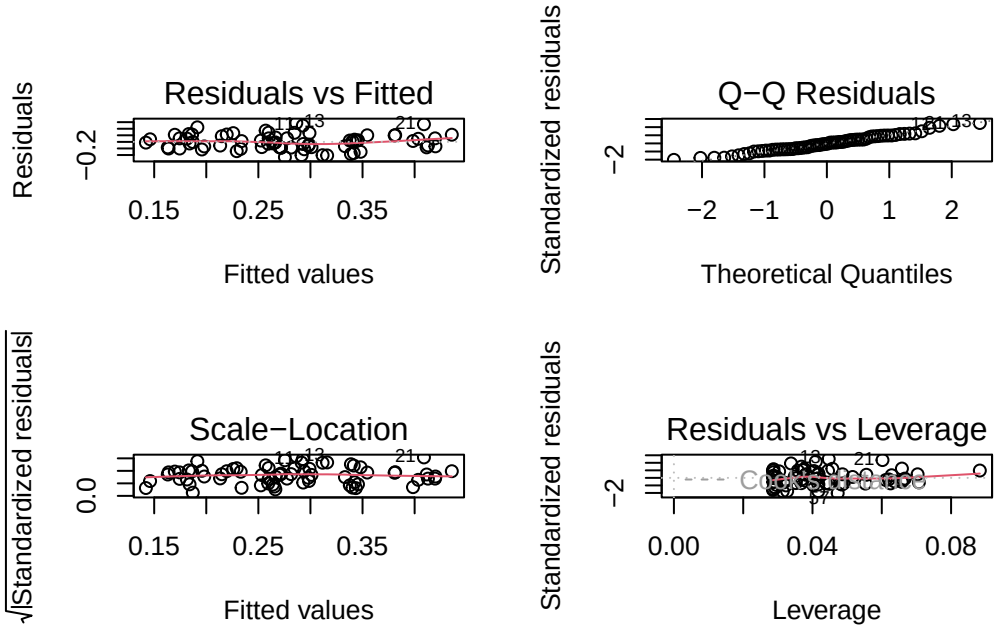
1. total biomass as a function of SLA, water treatment, and species

```
# saturated model
model1 <- lm(total_g ~ sla + water_treatment + species_name, #formula
             data = drought_exp_clean) #data frame
#viewing diagnostic plots in a 2X2 table
par(mfrow = c(2, 2))
#view model 1 diagnostic plots
plot(model1)
```



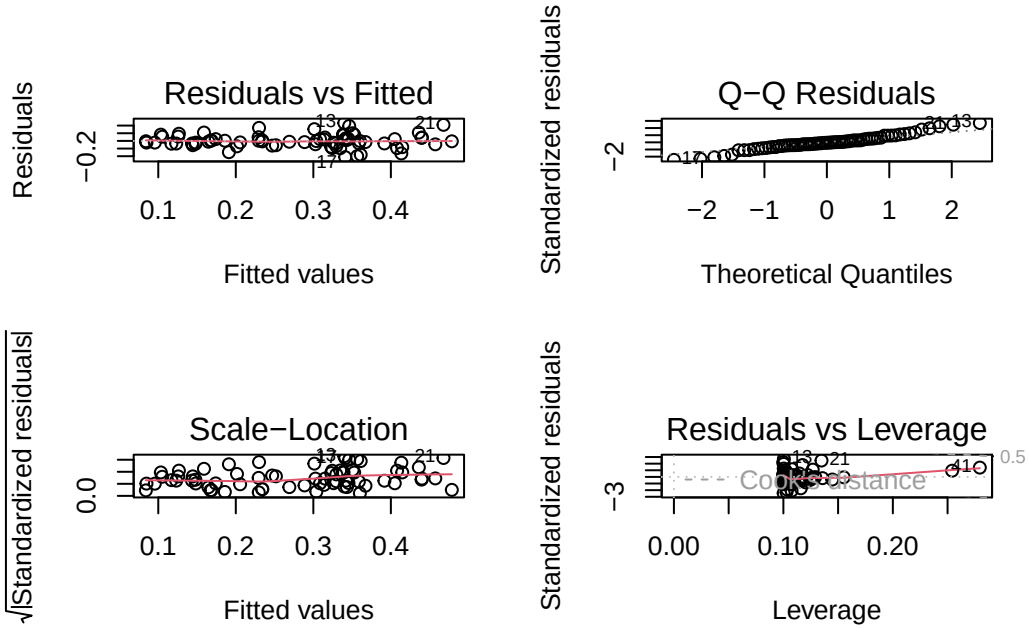
2. total biomass as a function of SLA and water treatment

```
model2 <- lm(total_g ~ sla + water_treatment, #formula
             data = drought_exp_clean) #data
#viewing diagnostic plots in a 2X2 table
par(mfrow = c(2, 2))
#view model 4 diagnostic plots
plot(model2)
```



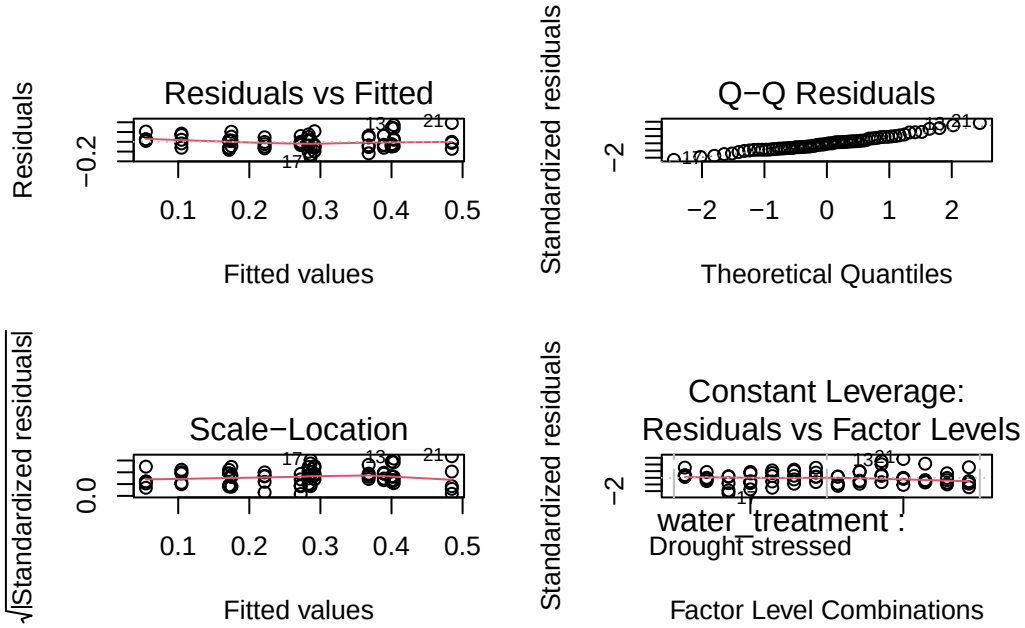
3. total biomass as a function of SLA and species

```
model3 <- lm(total_g ~ sla + species_name, #formula
             data = drought_exp_clean) #data framr
#viewing diagnostic plots in a 2X2 table
par(mfrow = c(2, 2))
#view model 3 diagnostic plots
plot(model3)
```



4. total biomass as a function of water treatment and species

```
model4 <- lm(total_g ~ water_treatment + species_name, #formula
             data = drought_exp_clean) #data
#viewing diagnostic plots in a 2X2 table
par(mfrow = c(2, 2))
#view model 4 diagnostic plots
plot(model4)
```




```
#creating model predictions data frame from model 4
model_preds_4 <- ggpredict(model4,
#basing predictions off of water treatment and species name
                           terms = c("water_treatment",
                                     "species_name"))
#viwing model 4 predictions as data frame
view(model_preds_4)
```

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

Problem 1a: 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)

```
#making a list of all models
models <- list(model0, model1, model2, model3, model4)

# Extract AIC values
aic_values <- sapply(models, AIC)
```

```
# Calculate delta AIC (difference from the minimum AIC)
delta_aic <- aic_values - min(aic_values)

# Create a a data frame with rows of model,predicator, AIC, and Delta AIC
summary_table <- data.frame(
#creating a column "model" with following categories
  Model = c("null", "model 1", "model 2", "model 3", "model 4"),
  #creating a column "predictors" with following categories
  Predictors = c("none", "specific leaf area, water treatment, and plant species", "specif
  #adding AIC values to data frame
  AIC = aic_values,
  #adding delta aic values to data frame
  Delta_AIC = delta_aic
)
#viwing data frame
view(summary_table)

final_table <- flextable(summary_table)

# Autofit columns
final_table <- flextable::autofit(final_table)

# Set table width and layout
final_table <- flextable::set_table_properties(final_table, width = 1, layout = "autofit")

# Adjust column widths of table
final_table <- flextable::set_table_properties(final_table, width = 1)

(knitr::opts_chunk$set(ft.shadow = FALSE))
```

```
$ft.shadow
[1] FALSE
```

```
# Print flextable
final_table
```

Model	Predictors	AIC	Delta_AIC
null	none	-75.15946	84.036493

Model	Predictors	AIC	Delta_AIC
model 1	specific leaf area, water treatment, and plant species	-157.48243	1.713524
model 2	specific leaf area and water treatment	-96.44059	62.755359
model 3	specific leaf and plant species	-127.07569	32.120263
model 4	water treatment and plant species	-159.19595	0.000000

Table 1. Linear Model Summary. The table above shows the listed linear models and associated predictor values used for their construction.The column header "Model" assigns a specific name to each linear model, while the "Predictor Variable" header outlines each specific predictor variable used for model construction.The AIC column provides a qualitative score to each model based on their level of complexity and predictive capacity, with lower scores indicating better performance. The Delta AIC column subtracts the lowest AIC score from each model's AIC score to indicate relative model performance.

Problem 1b: Write a 5-6 sentence "statistical methods" section. (8 points)

To examine the influence of species (categorical) and water treatment (categorical) on total biomass, I constructed five individual multiple linear models testing various combinations of these predictor variables (refer to Problem 1a). To determine the model that best described the influence of the listed predictor variables on total biomass, I first evaluated each model's individual AIC score along with its delta AIC score to select a model that balanced complexity and interpretability. The null model helped me to assess whether the inclusion of the predictor variables significantly improved model fit. The saturated model, provided a reference point for which predicator variables had significance in relation to the response variable (total biomass).

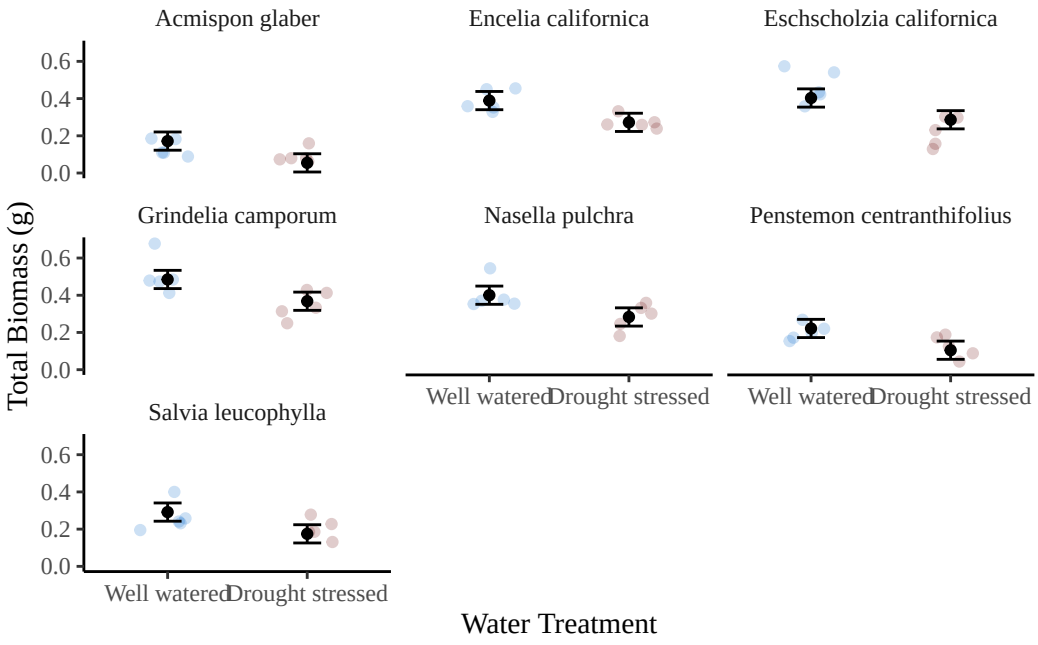
Next, I checked which model adhered best to the assumptions of a linear model (residuals are homoscedastic, residuals are normally distributed, outliers affecting the final model) to move forward with my final decision. To evaluate linear model assumptions visually, I checked for constant variance of residuals by looking for a straight line and evenly distributed residual points on the Residuals versus Fitted and Scale-Location plots, ensured that the residuals followed the reference line shown on the QQ Residuals plot, and confirmed that no outliers fell outside the Cook's distance line on the Residuals versus Leverage plot.

Problem 1c. Make a visualization of the model predictions with underlying data for your "best" model. (20 points)

```
# creating new data frame of model predictions for plotting
model4_preds_for_plotting <- model_preds_4 %>%
#renaming columns for easier data wrangling
  rename(water_treatment = x, #renamed to water treatment
         species_name = group) #rename to species_name

ggplot() +
  #jittering underlying data
   geom_jitter(data = drought_exp_clean,
    #setting x-axis
              aes(x = water_treatment,
    #setting y axis
                  y = total_g,
    #grouping up species name
                  group = species_name,
    #setting color to water treatment
                  color = water_treatment),
    #setting transparency of data
              alpha = 0.2,
    #bringing points cloer together
              width = 0.2) +
  #constructing the 95% confidence interval
  geom_errorbar(data = model4_preds_for_plotting,
              #setting x-axic
                aes(x = water_treatment,
              #setting ymin value
                    ymin = conf.low,
              #setting ymax value
                    ymax = conf.high),
              #setting width of error bars
                width = 0.2) +
  #plotting predicted value
  geom_point(data = model4_preds_for_plotting,
  #setting x-axis
             aes(x = water_treatment,
  #setting y-axis
                 y = predicted)) +
  #assigning colors to water treatment
```

```r
scale_color_manual(values = c("Drought stressed" = "#660000", "Well watered" = "#0066CC"
#faceting by species
facet_wrap(~species_name) +
#apply different theme
theme_classic() +
#getting rid of gridlines
theme(panel.grid = element_blank(),
      strip.background = element_blank(),
      #remove legend
      legend.position = "none",
      #changing front
      text = element_text(family = "Times")) +
#adding meaningful labels to graph
labs(x = "Water Treatment",
     y = "Total Biomass (g)")
```



Problem 1d: Write a caption for your visualization. (6 points)

Figure 1. Predictive modeling showing the influence of water treatments and plant species on total plant biomass (g). Points represent the predicted biomass of each individual plant species and their designated water treatment, calculated

using linear regression, with standard error bars showing the 95% confidence interval. The underlying data points represent individual data points for well-watered and drought-stressed treatment groups for each plant species. Data from Valliere, Justin; Zhang, Jacqueline; Sharifi, M.; Rundel, Philip (2019). Can we condition native plants to increase drought tolerance and improve restoration success? [Dataset]. Dryad. https://doi.org/10.5061/dryad.v0861f7

Problem 1e:Write a 3-4 sentence results section. (10 points)

```
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
```

```
#summarizing model prediction data for water treatment
water <- model4_preds_for_plotting %>%
  #group by water treatment
  group_by(water_treatment) %>%
```

```
  #calculating averages for confidence intervals and predicated value
  summarise(con_low = mean(conf.low),
            con_high = mean(conf.high),
            average_g = mean(predicted))
#summarizing model prediction data for species
species <- model4_preds_for_plotting %>%
  #group by species name
  group_by(species_name) %>%
  #calculating averages for confidence intervals and predicated value
  summarise(con_low = mean(conf.low),
            con_high = mean(conf.high),
            average_g = mean(predicted))


# Print the results of summary finding for water treatment
print(water)
```

```
# A tibble: 2 x 4
  water_treatment  con_low con_high average_g
  <fct>              <dbl>    <dbl>     <dbl>
1 Well watered       0.288    0.386     0.337
2 Drought stressed   0.172    0.270     0.221
```

```
# Print the results of summary finding for species
print(species)
```

```
# A tibble: 7 x 4
  species_name               con_low con_high average_g
  <fct>                        <dbl>    <dbl>     <dbl>
1 Encelia californica          0.282    0.380     0.331
2 Eschscholzia californica     0.296    0.394     0.345
3 Grindelia camporum           0.377    0.475     0.426
4 Acmispon glaber              0.0640   0.162     0.113
5 Nasella pulchra              0.293    0.391     0.342
6 Penstemon centranthifolius   0.114    0.212     0.163
7 Salvia leucophylla           0.184    0.282     0.233
```

The predictor variables of species type and water treatment (well watered and drought stressed) significantly predicted total plant biomass in grams ($F(7,62) = 27.08$, $p < 0.01$, $\alpha = 0.05$, adjusted $R^2 = 0.75$). On average well watered plants

had a higher biomass of 0.337 g ($p < 0.001$, 95% confidence interval = [0.2884552, 0.3864648]), while drought stressed plants had a average biomass of 0.220 ($p < 0.001$, 95% confidence interval = [0.1715095, 0.2695191]). On average, Grindelia camporum plant species had the highest plant biomass of 0.42637 ($p < 0.001$, 95% confidence interval = [0.37736522, 0.4753748]) followed by Eschscholzia californica with a biomass of 0.345 g ($p < 0.001$, 95% confidence interval = [0.28175522, 0.3797648]). While Acmispon glaber had the lowest plant biomass of 0.113 g (95% confidence interval = [0.06401522, 0.1620248]) follwed by the Penstemon centranthifolius plant species with a biomass of 0.163 g ($p > 0.005$, 95% confidence interval = [0.11404522, 0.2120548]).

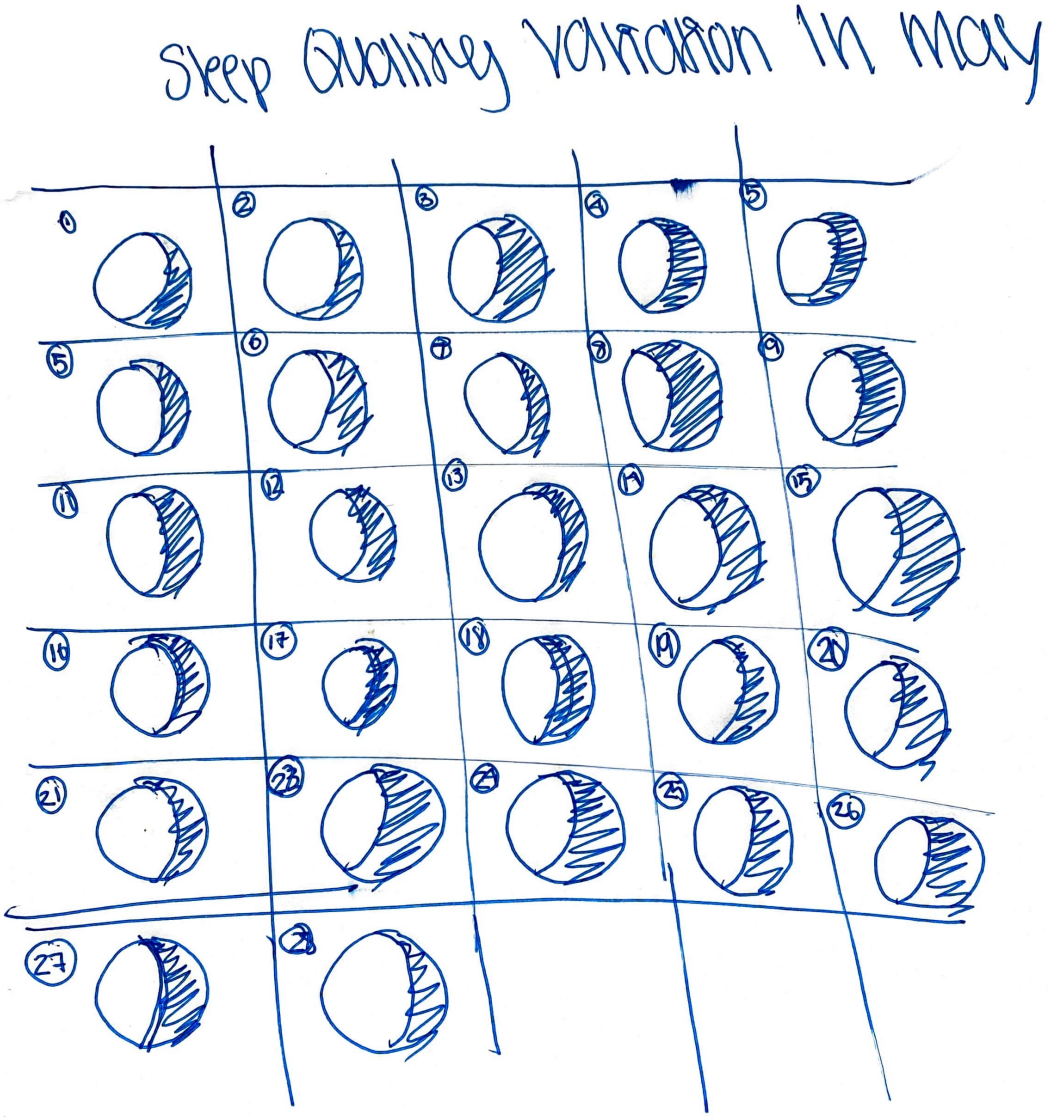
Problem 2. Affective visualization (24 points)

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

My personal data collection revolves around tracking sleep data, including the number of hours slept, time woken up, screen time before bed, etc. Affective visualization should thematically represent common symbols associated with sleep and nighttime to communicate my data collection to others. From this, I started to think about the moon and its various phases and how I could use them to represent some aspect of my data. Given this, each moon phase could represent an associated value of my sleep quality for the night.

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

```
#setting on the page
knitr::include_graphics("hw_figures/sleep_quality.pdf")
```


Sleep Quality variation in May

Probelm 2c: Make a draft of your visualization. (12 points)

```
#setting on the page
knitr::include_graphics("hw_figures/sleep_visualization.pdf")
```

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
	1 	2 	3 	4 	5 	6 
7 	8 	9 	10 	11 	12 	13 
14 	15 	16 	17 	18 	19 	20 
21 	22 	23 	24 	25 	26 	27 
28 	29	30	31			

Problem 2d: Write an artist statement. (8 points)

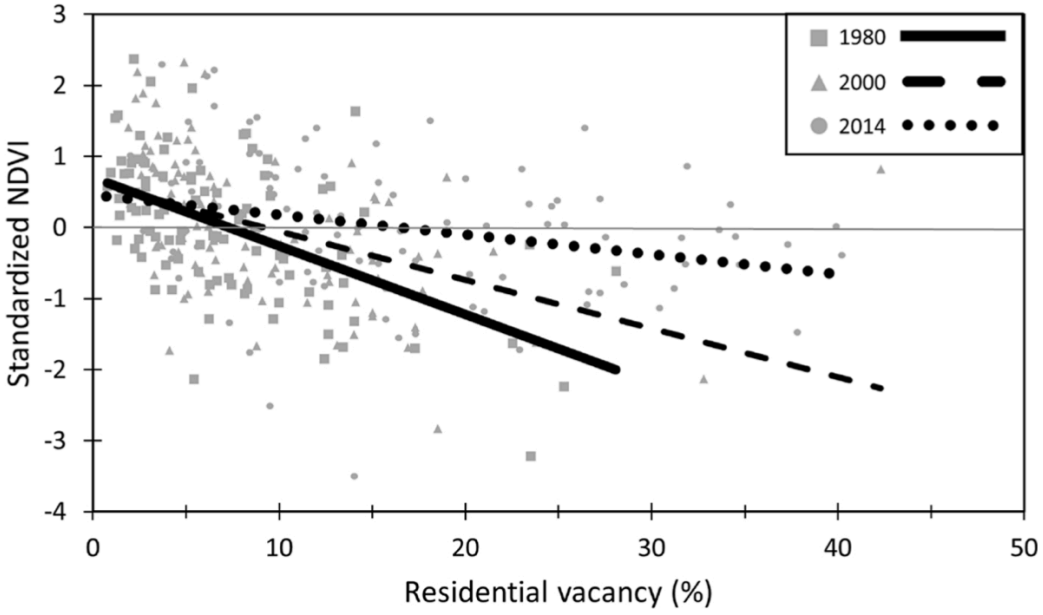
In the piece I am showing how my sleep quality varies over the month of month of May. As the moon wanes it is indicative of a lower quality of sleeps, while as it waxes it represents my sleep quality increasing. I knew that I want to draw upon figures representative to sleep and night time hence the decision to use the moon, however for formatting I was inspired my some of the artistic pieces presented in Stefanie Posavec and Giorgia Lupi's Dear Data project. The form of my work is a visual display I made in Canvas. To make this visual I found different images of the moon online to represent different sleep levels, and then inserted them into the calendar I made to play on the idea of time and changes in sleep quality throghout the month of May.

Problem 3. Statistical critique (36 points)

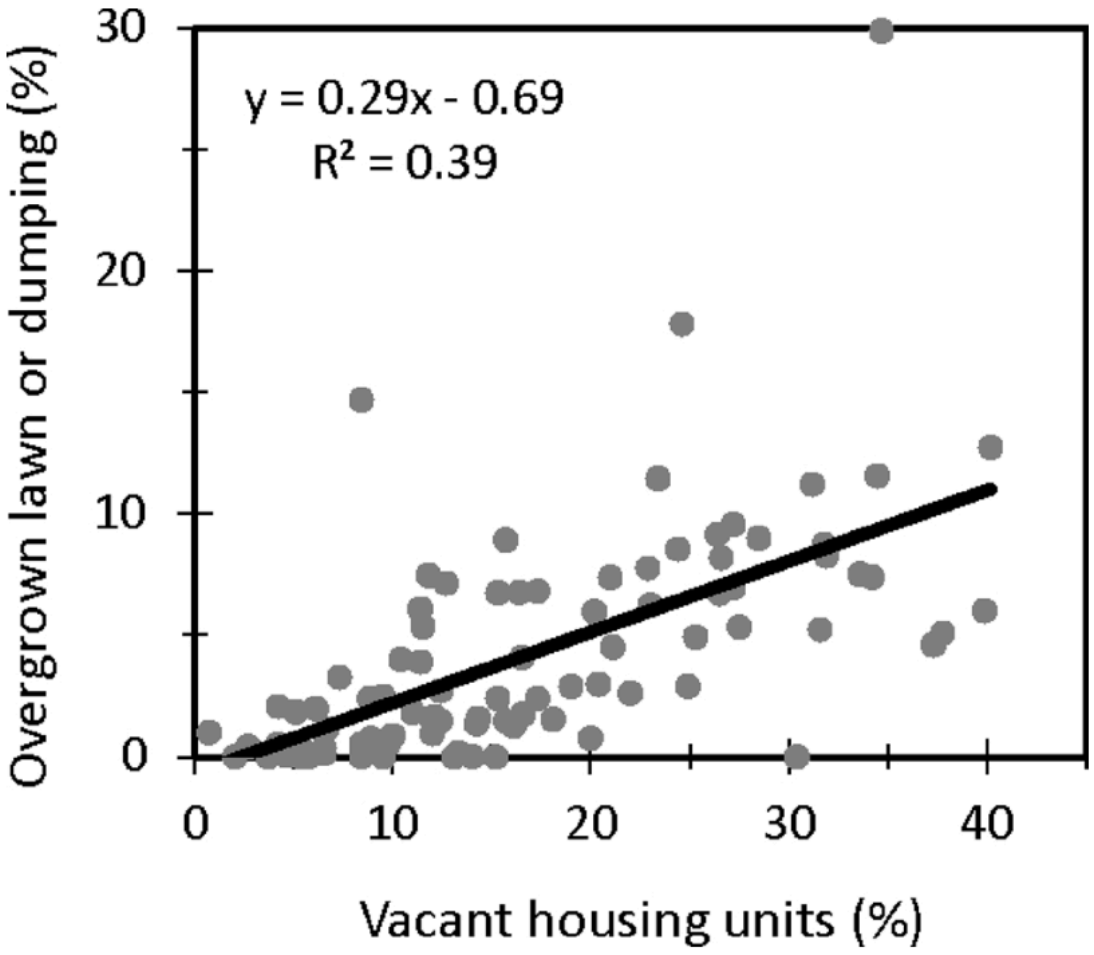
Problem 3a: Revisit and summarize (6 points)

The authors used linear regression models to examine the relationship between housing vacancy and urban greening in Toledo, Ohio. They constructed separate linear regression models for the years 1980, 2000, and 2004 to evaluate the relationship between mean NDVI values and vacancy rates in Toledo. Additionally, they created a linear regression model to examine the influence of vacant housing units on the amount of overgrown lawns and dumping (a sign of blight). Lastly, the authors calculated Spearman correlations to explore the relationship between vacancy and race, wealth, poverty, and educational attainment.

```
#setting on the page
knitr::include_graphics("hw_figures/Figure_3.pdf")
```


1980
2000
2014
Standardized NDVI
Residential vacancy (%)
3
2
1
0
-1
-2
-3
-4
0
10
20
30
40
50

```
#setting image on the page
knitr::include_graphics("hw_figures/Figure_4.pdf")
```


y = 0.29x - 0.69
R² = 0.39
Overgrown lawn or dumping (%)
Vacant housing units (%)
0
10
20
30
40

```r
#setting image on the page
knitr::include_graphics("hw_figures/Table_2.pdf")
```

Table 2
Spearman correlation coefficients between vacancy rate and select population variables in Toledo. For all correlations, $p < 0.001$.

	1980	2000	2014
Median household income	−0.73	−0.84	−0.62
Median home value	−0.63	−0.70	−0.68
White population	−0.75	−0.78	−0.66
High school graduates	−0.61	−0.78	−0.69
Poverty rate	0.77	0.83	0.64

Problem 3b: Visual clarity (10 points)

In Figures 3 and 4, the authors used meaningful x and y axes to signify their predictor and response variables for readers. For both of these linear regression models, the authors showcased the underlying data alongside the model predictions. In Figure 3, the authors differentiated data points for three grouping variables (1980, 2000, and 2014) by assigning each year and its associated data unique symbols and shapes. This helped to distinguish between the different years, boosting the level of interpretation for the graph.

Problem 3c: Aesthetic clarity (10 points)

The authors focused on minimizing unnecessary visual elements required for interpreting the linear models. In both figures, visual elements are restricted to the model prediction line and the associated underlying data, enhancing interpretation while reducing complexity. One visual element that could be removed is the legend in Figure 4, faceting by each year would enable conveying the same information while reducing visual clutter. Overall the data:ink ratio in the figures is relatively strong with few areas for improvement.

Problem 3d: Recommendations (can be longer than 4 sentences, 10 points)

For both Figures 3 and 4, the opacity and lack of color of the underlying data make it difficult to distinguish between points. This design choice causes individual points to merge together. Given this, I recommend that the authors reduce the opacity of the individual points and assign colors to increase clarity. Additionally, throughout the class, we have emphasized the importance of showing uncertainty in all visualizations for greater transparency in data communication. Therefore, for the linear regression models in Figures 3 and 4, I suggest including a 95 percent confidence interval around the model prediction to enhance data transparency. Lastly, as mentioned before, while the data-to-ink ratio is relatively good in the figures, there is room for improvement. In Figure 4, the authors should facet the data by year, which would eliminate the need for a legend, further reducing visual clutter.