ES 193DS Homework 3

Ella Stookey

2024-06-02

Forked Repository: ADD LINK

Preparations

Reading in packages

```
# hide messages and warnings
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)

suppressPackageStartupMessages({
  # 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(knitr)
})

drought_exp <- read_xlsx(path = here("data",
                                     "Valliere_etal_EcoApps_Data.xlsx"),
                         sheet = "First Harvest")

# quick look at data
str(drought_exp)
```

```
tibble [70 x 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 0.0
 $ 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

```
# 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
```

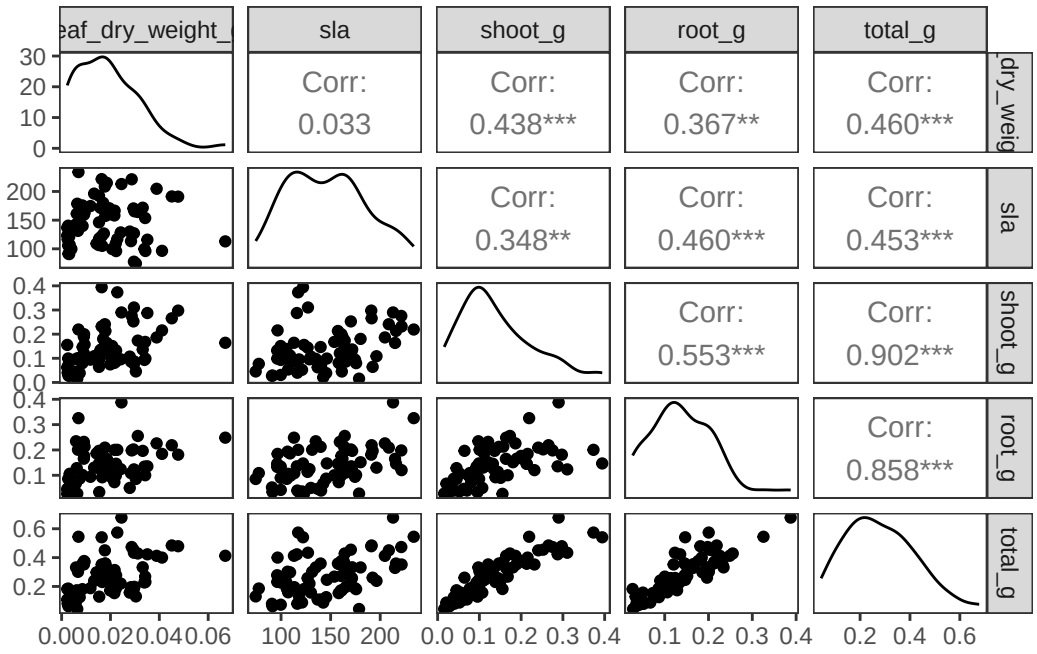
```
    species == "GRICAM" ~ "Grindelia camporum", # great valley gumweed
    species == "SALLEU" ~ "Salvia leucophylla", # Purple sage
    species == "STIPUL" ~ "Nasella pulchra", # Purple needlegrass
    species == "LOTSCO" ~ "Acmispon glaber" # deerweed
  )) %>%
  relocate(species_name, .after = 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

drought_exp_clean
```

```
# A tibble: 70 x 15
   species species_name   water water_treatment rep_number height_cm leaf_number
   <chr>   <chr>          <chr> <chr>                <dbl>     <dbl>       <dbl>
 1 ENCCAL  Encelia calif~ WW    Well watered             1       5.8          11
 2 ENCCAL  Encelia calif~ WW    Well watered             2       4.9           8
 3 ENCCAL  Encelia calif~ WW    Well watered             3       8.4          11
 4 ENCCAL  Encelia calif~ WW    Well watered             4       6.5          12
 5 ENCCAL  Encelia calif~ WW    Well watered             5       7.1          10
 6 ENCCAL  Encelia calif~ DS    Drought stress~          1       3.2           7
 7 ENCCAL  Encelia calif~ DS    Drought stress~          2       4.4           7
 8 ENCCAL  Encelia calif~ DS    Drought stress~          3       4.2          10
 9 ENCCAL  Encelia calif~ DS    Drought stress~          4       4.5           8
10 ENCCAL  Encelia calif~ DS    Drought stress~          5       3.9           6
# i 60 more rows
# i 8 more variables: leaf_dry_weight_g <dbl>, leaf_area_cm2 <dbl>, sla <dbl>,
#   total_la <dbl>, shoot_g <dbl>, root_g <dbl>, total_g <dbl>, r_s <dbl>
```

correlations

```
ggpairs(drought_exp_clean, # data frame
        columns = c("leaf_dry_weight_g", # columns to visualize
                    "sla",
                    "shoot_g",
                    "root_g",
                    "total_g"),
        upper = list(method = "pearson")) + # calculating Pearson correlation coefficient
```

```
  theme_bw() + # cleaner theme
  theme(panel.grid = element_blank()) # getting rid of gridlines
```




```
# bottom left scatterplots of listed variables -- Leaf dry weight on x axis, y axis is total
# upper right shows Pearson's correlation -- positively correlated
```

Problem 1. Multiple linear regression: model selection and construction

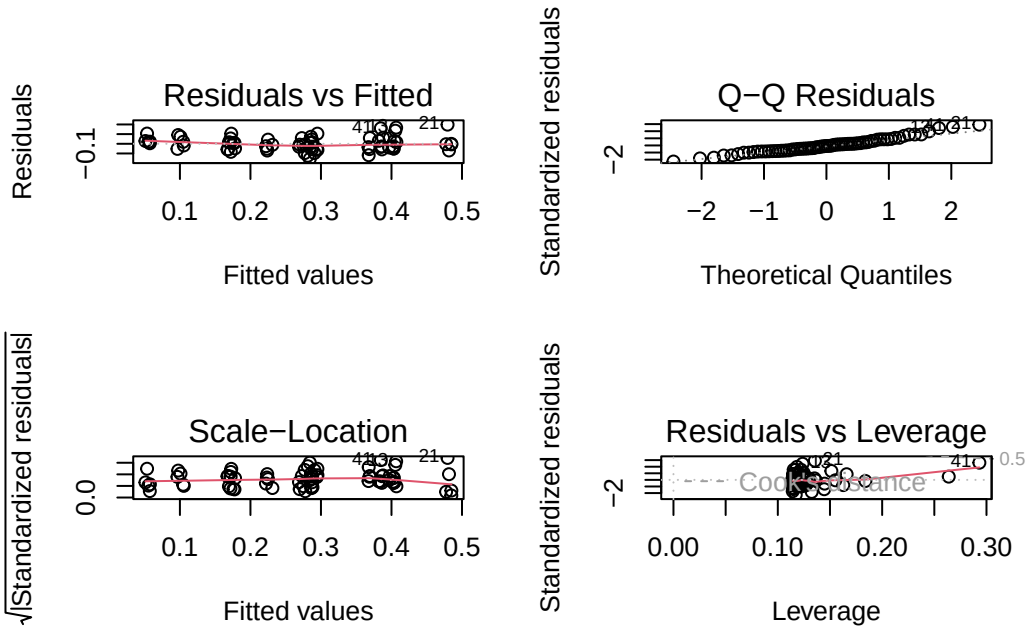
Part a

0. Null model

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

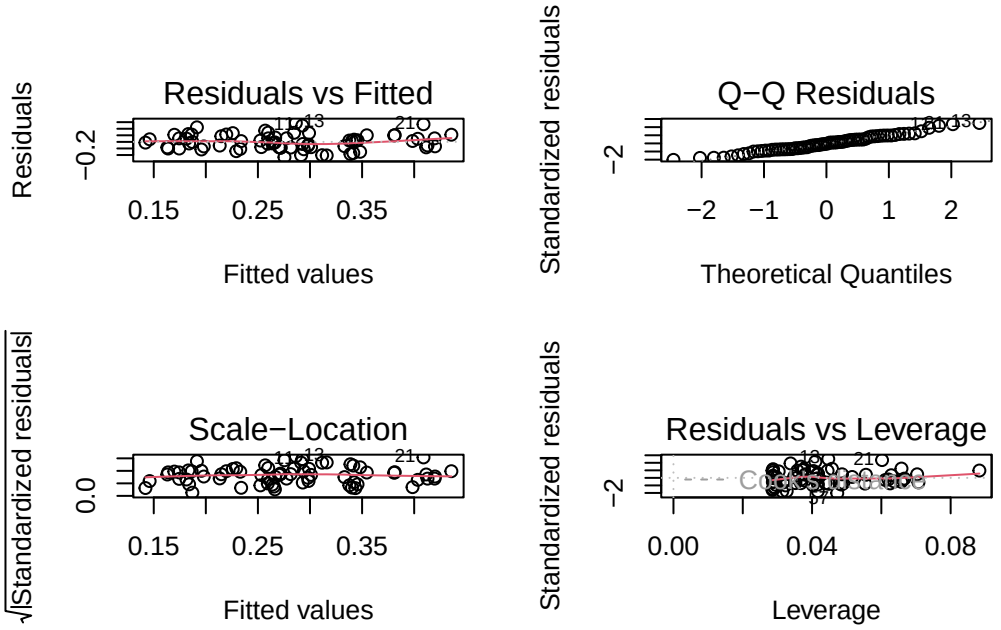
1. total biomass as a function of SLA, water treatment, and species

```
# saturated model
model1 <- lm(total_g ~ sla + water_treatment + species_name,
             data = drought_exp_clean)

par(mfrow = c(2, 2))
plot(model1)
```



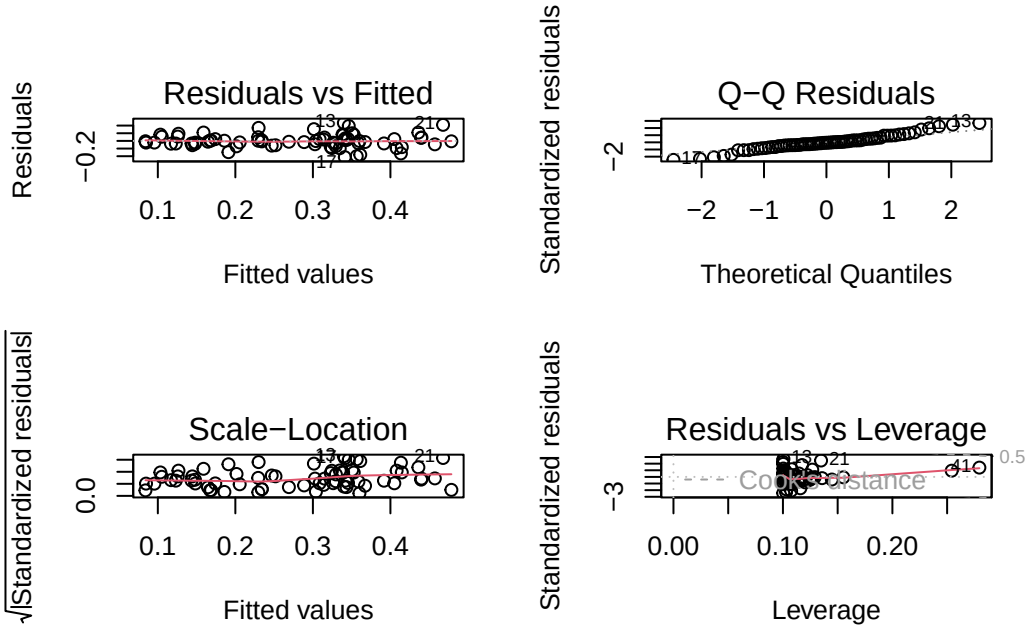
2. total biomass as a function of SLA and water treatment

```
model2 <- lm(total_g ~ sla + water_treatment,
             data = drought_exp_clean)

par(mfrow = c(2, 2))
plot(model2)
```



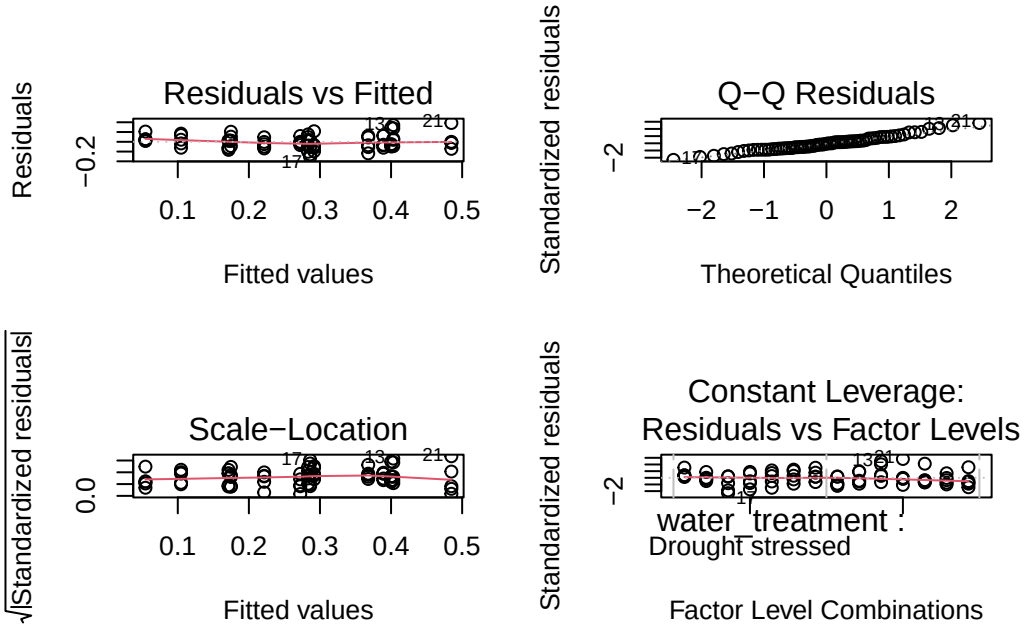
3. total biomass as a function of SLA and species

```
model3 <- lm(total_g ~ sla + species_name,
             data = drought_exp_clean)

par(mfrow = c(2, 2))
plot(model3)
```



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))
plot(model4)
```




```
# Model selection table to see AIC and delta values

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)
```

```
# See model 4 coefficents
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
```

Table presentation

```r
model_info <- data.frame(
  Model_number = c("0 (null)", "1 (saturated)", 2, 3, 4),
  Predictors = c("None", "SLA, Treatment, Species", "SLA, Treatment", "SLA, Species","Treatme
)
```

```r
kable(model_info,
      caption = "Table 1",
      col.names = c("Model", "Predictors")) #label columns
```

Table 1: Table 1

Model	Predictors
0 (null)	None
1 (saturated)	SLA, Treatment, Species
2	SLA, Treatment

Model	Predictors
3	SLA, Species
4	Treatment, Species

Part b

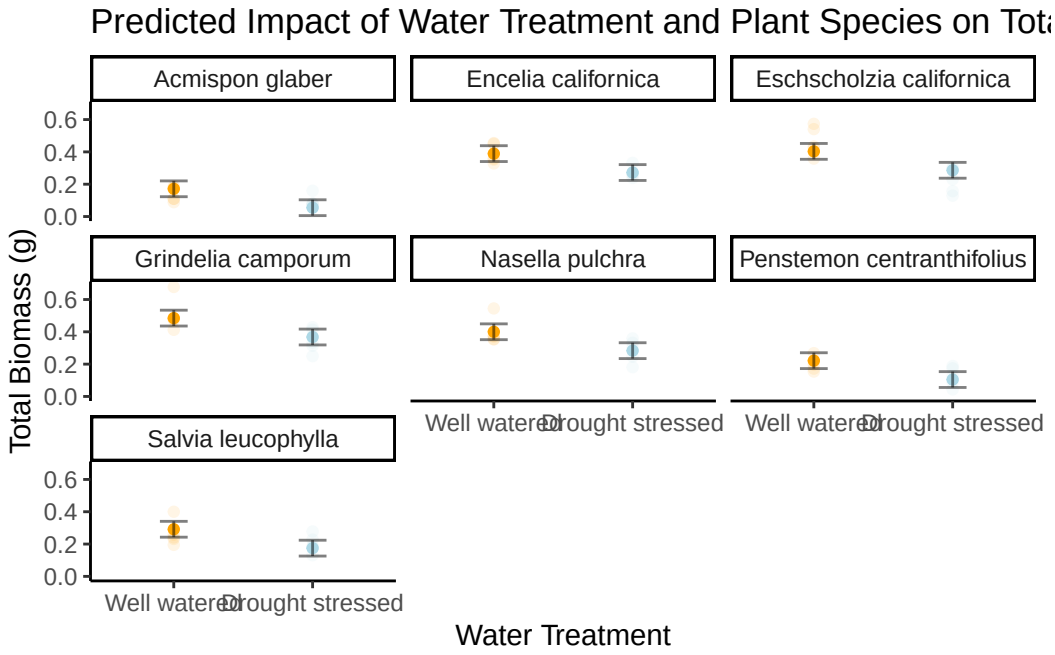
To examine the influence of specific leaf area, water treatment, and plant species type on total plant biomass, I have constructed and analyzed five linear regression models. To determine the model that best described the impact these potential influences have on plant biomass, a model selection table was used. This revealed model 4, which examined the water treatment and plant species variables, to have the lowest Akaike Information Criterion (AIC) of -156.2 and a delta of 0. These values suggest that model 4 is best because it is descriptive, but not too complex. To confirm this, I had to ensure that model 4 conformed to the assumptions of a linear model by examining its diagnostic plots. Beginning with the homeoscadatic models, the residuals are scattered along a straight line and have an even distribution above and below the red line indicating the constant variance. The Q-Q plot tests if the data is normally distributed, which it appears to be as it follows a linear path. The last plot shows no outliers with significant influence because if there were, there would be a red dashed line with points falling outside of it. Overall, the diagnostic plots confirm the model selection tables' suggestion that model 4 is best. Lastly, model 4's coefficient summary was examined to further solidify the conclusion that it is the best model. After looking at the slope, intercept, and level estimates, model 4 was confirmed best as they showed the significant differences between water treatments and the varying biomasses across plant species.

Part c

Model predictions

```
model4_preds <- ggpredict(model4,
                          terms = c(
                                    "water_treatment",
                                    "species_name"))

# use View(model_preds) to see the predictions as a data frame
# use model_preds to see the predictions formatted nicely
```

```r
# creating new data frame of model predictions for plotting
model_preds_for_plotting <- model4_preds %>%
  rename(water_treatment = x, # renaming columns to make this easier to use
         species_name = group)

# use View(model_preds_for_plotting)
# to compare this to the original model_preds data frame

ggplot() +
  # underlying data
  geom_point(data = drought_exp_clean,
             alpha = 0.1,
             aes(x = water_treatment,
                 y = total_g,
                 color = water_treatment)) +
  # add prediction data
  geom_point(data = model_preds_for_plotting,
             aes(x = water_treatment,
                 y = predicted,
                 color = water_treatment)) +
  # model prediction 95% CI errorbar
  geom_errorbar(data = model_preds_for_plotting,
              aes(x = water_treatment,
                  y = predicted,
                  ymin = conf.low,
                  ymax = conf.high,
                  fill = water_treatment),
              width = 0.2, alpha = 0.5) +
  # cleaner theme
  theme_classic() +
  # creating different panels for species
  facet_wrap(~species_name) +
  theme(legend.position = "none") + # removed legend
  labs(title = "Predicted Impact of Water Treatment and Plant Species on Total Biomass",
       x = "Water Treatment",
       y = "Total Biomass (g)") + # change plot title and axes labels
  scale_color_manual(values = c("Well watered" = "orange",
                                "Drought stressed" = "lightblue")) # change colors
```



Part d

Part e

Model 4, the best model and shown above in the figure, has the predictors of water treatment and plant species, both of which best describe total mass. Model 4 has the lowest AIC value (AIC = -156.2) of all the models, a delta of 0, an F-statistic of 27.08, and a p-value of 2.2e-16, all of which support that this is a significant model. On average between water treatments, those that were well watered tended to have a higher total biomass than those that were drought stressed.

Problem 2. Affective visualization

Part a

For my personal data set, where I am examining the distance traveled each day, I could use a bar graph and outline the perimeter of each peak. In doing so, the graph will appear to be "hilly". Since my data is about driving, I will turn this into a scene with a car driving over hills (ie the bar graph).

Part b

Part c

Part d

For my visualization, I have created a scene of a car traveling on a hilly road. The hills represent the distance traveled (in miles) each day, some days peaking while others are flatter. This work was done on a digital coloring platform called Notability. I began by importing a screenshot of my data (bar graph) and from there I traced and colored it. I finished it off by adding details, such as the road and cars. During this process I was unsure whether to keep each date and numerical value, however ultimately did decide to keep them because I thought it helped the viewer see the information clearer.

Problem 3. Statistical critique

Part a

To examine the long-term effects of a wildfire on soil nutrients and makeup, the researchers used a two-way ANOVA test and if significant differences were found ($p < 0.05$), a Tukey HSD post-hoc test was applied. The authors represented these statistical tests in three tables. Table 1 shows the results of the ANOVA test and Tables 2 & 3 show the descriptive statitics for certain nutrients. In addition to the tables, there were two figures. The first was a map that illustrated the study location and areas with varying fire severity. The second figure was an RDA for the relation between factors 1 and 2.

Part b

All three tables were very clear, with descriptive captions and column and row labels. Figure 1 was also simple to understand because it consisted of images and maps for context. However, figure 2 was significantly more confusing to understand because I have never looked at a redundancy analysis (RDA) plot before. There are no units on the x and y axis and at first glance the numbers seem quiet arbitrary. After researching how to read the plot, it made more sense and I could see how the variables' summary statistics (means and standard deviations) were being shown. No model predictions were in the matrix, but rather just the collected data.

Part c

The tables all hold a lot of information and data making them seem a bit visually cluttered. However, this was crucial information for the researchers to show so it was necessary to include it all. Figure 2, the RDA plot, had a very good data to ink ratio, only consisting of a few lines, two colors, and minimal lettering.

Part d

For the tables, I think they could have been made clearer if titles/ labels were bolded or a larger font. This would help differentiate the organizational aspects from the large amounts of data. Additionally having lines in the table would have helped section off different nutrients' information instead of it all blending together. As for the figure, I wish the caption included more information about what each percent on the sides of the plot and the axes represent. Although information about this is included in the text of the paper, I think including it on the figure is crucial so the reader can get a clear picture of the results from the experiment.