Assignment 6: GLMs week 1 (t-test and ANOVA)

Thomas Hancock

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on t-tests and ANOVAs.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Salk_A06_GLMs_Week1.Rmd") prior to submission.

The completed exercise is due on Tuesday, February 18 at 1:00 pm.

Set up your session

1. Check your working directory, load the `tidyverse`, `cowplot`, and `agricolae` packages, and import the NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv dataset.

2. Change the date column to a date format. Call up `head` of this column to verify.

```
#1 Load packages and input data
#install.packages("agricolae")
library(tidyverse)
library(cowplot)
library(agricolae)
getwd()
```

```
## [1] "C:/Users/thoma/Thomas/2018 Grad School/Duke MEM/ENV 872/Environmental_Data_Analytics_2020"
```

```
NTL.PP.Nutrients.Processed <-
  read.csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv")



#2 Set date format
NTL.PP.Nutrients.Processed$sampledate <- as.Date(NTL.PP.Nutrients.Processed$sampledate,
                                                 format = "%Y-%m-%d")
head(NTL.PP.Nutrients.Processed$sampledate) # Check the first 6 values
```

```
## [1] "1991-05-20" "1991-05-20" "1991-05-20" "1991-05-20" "1991-05-20"
## [6] "1991-05-20"
```

Wrangle your data

3. Wrangle your dataset so that it contains only surface depths and only the years 1993-1996, inclusive. Set month as a factor.

```
NTL.PP.Nut.Filtered <- filter(NTL.PP.Nutrients.Processed, year4 >= 1993 & year4 <= 1996,
                              depth == 0.00) # Filter data for surface records in date range

NTL.PP.Nut.Filtered$month <- as.factor(NTL.PP.Nut.Filtered$month) # Set month as a factor
```

Analysis

Peter Lake was manipulated with additions of nitrogen and phosphorus over the years 1993-1996 in an effort to assess the impacts of eutrophication in lakes. You are tasked with finding out if nutrients are significantly higher in Peter Lake than Paul Lake, and if these potential differences in nutrients vary seasonally (use month as a factor to represent seasonality). Run two separate tests for TN and TP.

4. Which application of the GLM will you use (t-test, one-way ANOVA, two-way ANOVA with main effects, or two-way ANOVA with interaction effects)? Justify your choice.

   Answer: I will use two-way ANOVA with interaction effects (months*lakename) because we are looking at whether the values vary by both lake and month (two factors), and we want to know if the differences between the lakes are affected by the month (interaction effect).

5. Run your test for TN. Include examination of groupings and consider interaction effects, if relevant.

6. Run your test for TP. Include examination of groupings and consider interaction effects, if relevant.

```
#5
tn.interaction <- with(NTL.PP.Nut.Filtered, interaction(lakename, month)) # Find interactions

pp.tn.anova.2way <- aov(data = NTL.PP.Nut.Filtered, tn_ug ~ tn.interaction) # 2 Way ANOVA

tn.groups <- HSD.test(pp.tn.anova.2way, "tn.interaction", group = TRUE) # Find stats groups
tn.groups # Display statistical groups/letters
```

```
## $statistics
##   MSerror Df     Mean       CV
##   67792.1 97 487.4077 53.41917
##
## $parameters
##    test         name.t ntr StudentizedRange alpha
##   Tukey tn.interaction  10         4.579991  0.05
##
## $means
##                  tn_ug       std  r     Min      Max      Q25      Q50      Q75
## Paul Lake.5  300.5115  67.85647  6 244.870  417.345 251.0738 275.0400 329.5267
## Paul Lake.6  324.1245 117.32193 17  45.670  439.984 307.8120 342.8260 422.2600
## Paul Lake.7  353.6341  40.78474 14 281.421  412.669 328.0188 351.6630 385.5945
## Paul Lake.8  336.5081 118.22435 14 163.148  499.251 233.8633 356.6185 423.1365
## Paul Lake.9  406.3360 169.15898  3 223.799  557.812 330.5980 437.3970 497.6045
## Peter Lake.5 384.9389  62.65797  7 312.133  460.791 333.7260 373.0810 440.5575
## Peter Lake.6 609.0427 379.99046 16 379.781 1962.902 462.9225 497.8530 606.3447
## Peter Lake.7 709.8848 422.31321 13 352.001 2048.151 571.0920 590.7920 707.7710
## Peter Lake.8 745.9833 349.34126 15 448.049 1924.631 579.3500 688.5110 781.0950
## Peter Lake.9 550.4680 183.97504  2 420.378  680.558 485.4230 550.4680 615.5130
##
## $comparison
## NULL
##
## $groups
##                  tn_ug groups
## Peter Lake.8 745.9833      a
## Peter Lake.7 709.8848      a
## Peter Lake.6 609.0427     ab
## Peter Lake.9 550.4680     ab
## Paul Lake.9  406.3360     ab
## Peter Lake.5 384.9389     ab
## Paul Lake.7  353.6341      b
## Paul Lake.8  336.5081      b
## Paul Lake.6  324.1245      b
## Paul Lake.5  300.5115      b
##
## attr(,"class")
## [1] "group"
```

```
#6
tp.interaction <- with(NTL.PP.Nut.Filtered, interaction(lakename, month)) # Find interactions

pp.tp.anova.2way <- aov(data = NTL.PP.Nut.Filtered, tp_ug ~ tp.interaction) # 2 Way ANOVA

tp.groups <- HSD.test(pp.tp.anova.2way, "tp.interaction", group = TRUE) # Find stats groups
tp.groups # Display statistical groups/letters
```

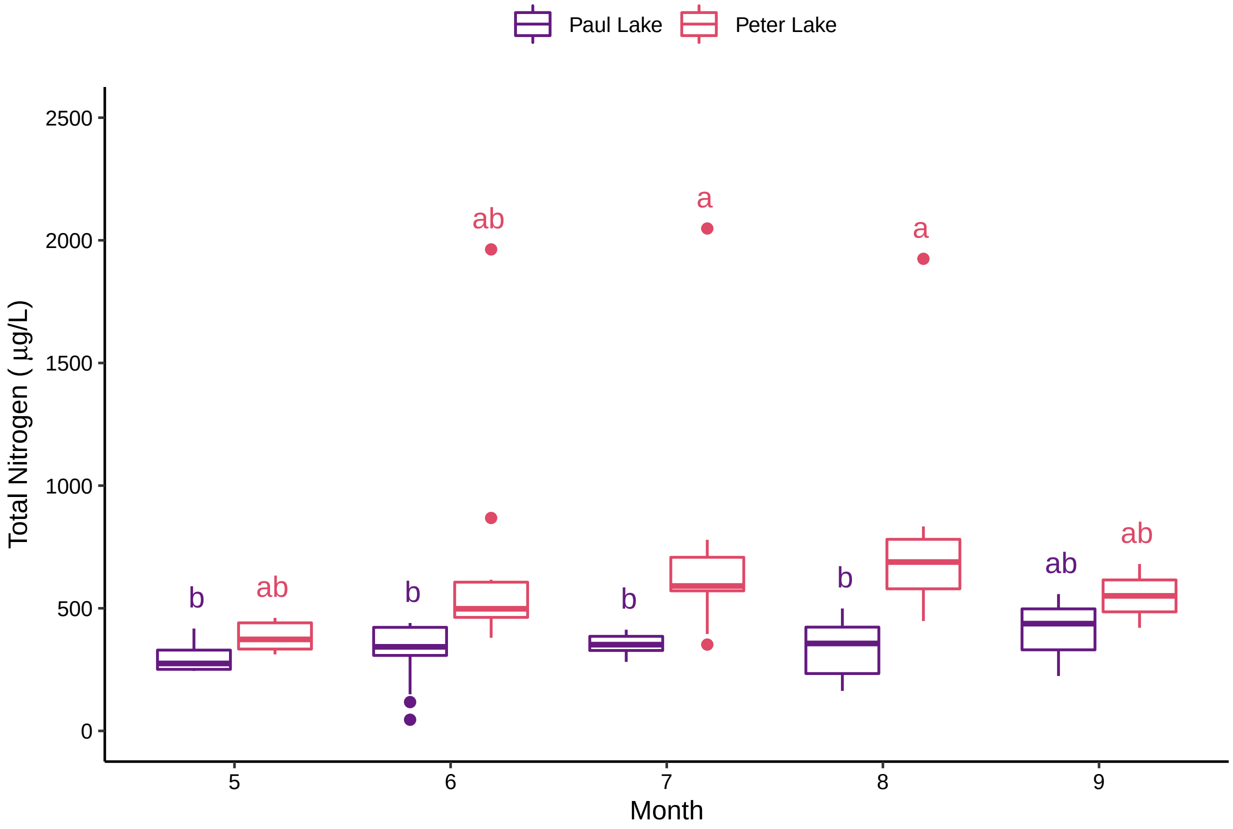
```
## $statistics
##    MSerror  Df     Mean      CV
##   103.4055 119 19.07347 53.3141
##
## $parameters
##    test         name.t ntr StudentizedRange alpha
##   Tukey tp.interaction  10         4.560262  0.05
##
## $means
##                  tp_ug       std  r    Min    Max     Q25     Q50      Q75
## Paul Lake.5  11.474000  3.928545  6  7.001 17.090  8.1395 11.8885 13.53675
## Paul Lake.6  10.556118  4.416821 17  1.222 16.697  7.4430 10.6050 13.94600
## Paul Lake.7   9.746889  3.525120 18  4.501 21.763  7.8065  9.1555 10.65700
## Paul Lake.8   9.386778  1.478062 18  5.879 11.542  8.4495  9.6090 10.45050
## Paul Lake.9  10.736000  3.615978  5  6.592 16.281  8.9440 10.1920 11.67100
## Peter Lake.5 15.787571  2.719954  7 10.887 18.922 14.8915 15.5730 17.67400
## Peter Lake.6 28.357889 15.588507 18 10.974 53.388 14.7790 24.6840 41.13000
## Peter Lake.7 34.404471 18.285568 17 19.149 66.893 21.6640 24.2070 50.54900
## Peter Lake.8 26.494000  9.829596 19 14.551 49.757 21.2425 23.2250 27.99350
## Peter Lake.9 26.219250 10.814803  4 16.281 41.145 19.6845 23.7255 30.26025
##
## $comparison
## NULL
##
## $groups
##                  tp_ug groups
## Peter Lake.7 34.404471      a
## Peter Lake.6 28.357889     ab
## Peter Lake.8 26.494000    abc
## Peter Lake.9 26.219250   abcd
## Peter Lake.5 15.787571    bcd
## Paul Lake.5  11.474000     cd
## Paul Lake.9  10.736000     cd
## Paul Lake.6  10.556118      d
## Paul Lake.7   9.746889      d
## Paul Lake.8   9.386778      d
##
## attr(,"class")
## [1] "group"
```

7. Create two plots, with TN (plot 1) or TP (plot 2) as the response variable and month and lake as the predictor variables. Hint: you may use some of the code you used for your visualization assignment. Assign groupings with letters, as determined from your tests. Adjust your axes, aesthetics, and color palettes in accordance with best data visualization practices.

8. Combine your plots with cowplot, with a common legend at the top and the two graphs stacked vertically. Your x axes should be formatted with the same breaks, such that you can remove the title and text of the top legend and retain just the bottom legend.

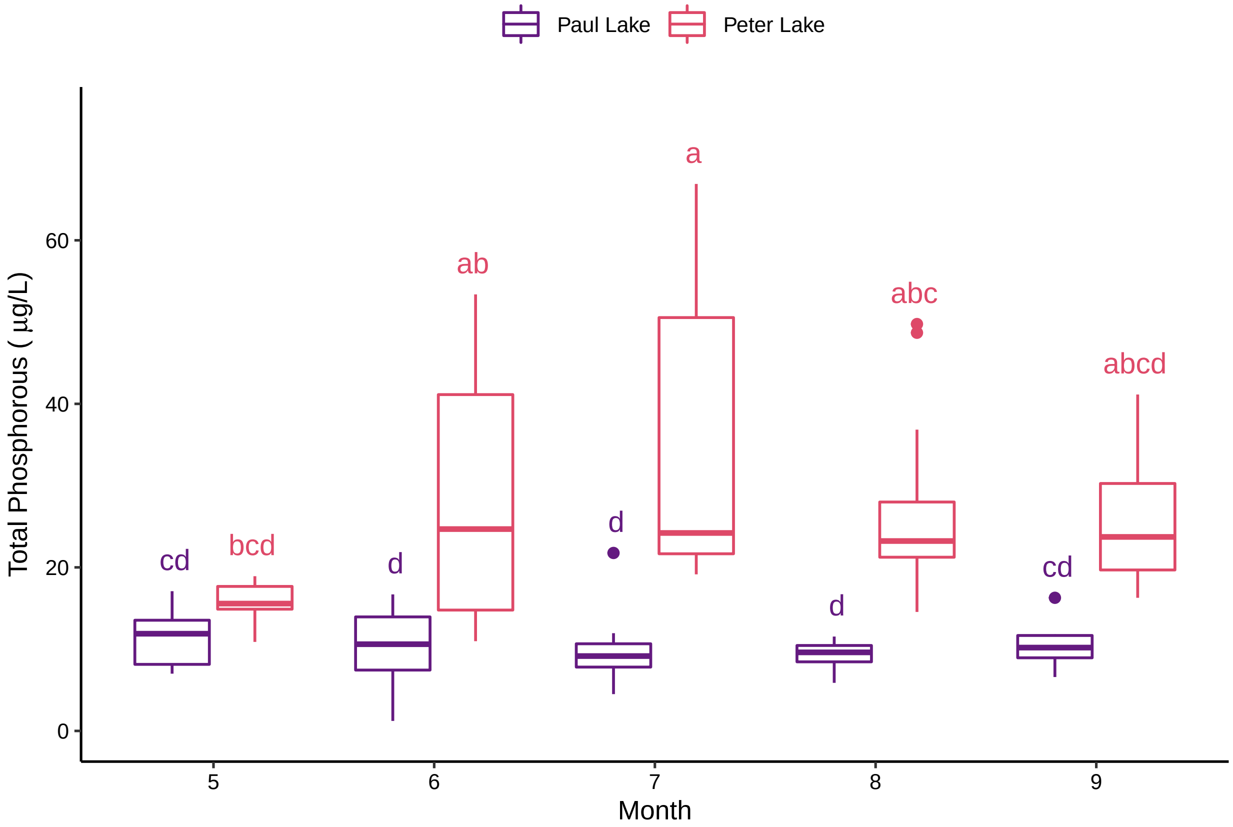
```
#7

myTheme <- theme_classic(base_size = 10) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top") # Define a theme based off of the classic theme

theme_set(myTheme) # Set defined theme to default

# Create a data frame of the statistical letters in the proper order
tn.letters <- tn.groups$groups[match(levels(tn.interaction), rownames(tn.groups$groups)),]

# Create plot for total nitrogen
tn.anova.plot <- ggplot(NTL.PP.Nut.Filtered, aes(x = month, y = tn_ug, color = lakename)) +
  geom_boxplot() +
  labs(x = "Month", y = expression(paste("Total Nitrogen ( ", mu, "g/L)")), color = "") +
  scale_color_viridis_d(option = "magma", begin = 0.3, end = 0.6) +
  coord_cartesian(ylim = c(0,2500)) + # Expand y-axis to include the stat letters
  stat_summary(geom = "text", fun.y = max, vjust = -1, size = 4, # Show stat letters
               position = position_dodge(0.7), show.legend = FALSE,
               label = c("ab", "b", "ab", "b", "a", "b", "a", "b", "ab", "ab"))

print(tn.anova.plot)
```

```
## Warning: Removed 23 rows containing non-finite values (stat_boxplot).

## Warning: Removed 23 rows containing non-finite values (stat_summary).
```



```r
# Create a data frame of the statistical letters in the proper order
tp.letters <- tp.groups$groups[match(levels(tp.interaction), rownames(tp.groups$groups)),]

# Create plot for total phosphorous
tp.anova.plot <- ggplot(NTL.PP.Nut.Filtered, aes(x = month, y = tp_ug, color = lakename)) +
  geom_boxplot() +
  labs(x = "Month", y = expression(paste("Total Phosphorous ( ", mu, "g/L)")), color = "") +
  scale_color_viridis_d(option = "magma", begin = 0.3, end = 0.6) +
  coord_cartesian(ylim = c(0,75)) + # Expand y-axis to include stat letters
  stat_summary(geom = "text", fun.y = max, vjust = -1, size = 4, # Show stat letters
               position = position_dodge(0.7), show.legend = FALSE,
               label = c("bcd", "cd", "ab", "d", "a", "d", "abc", "d", "abcd", "cd"))

print(tp.anova.plot)
```

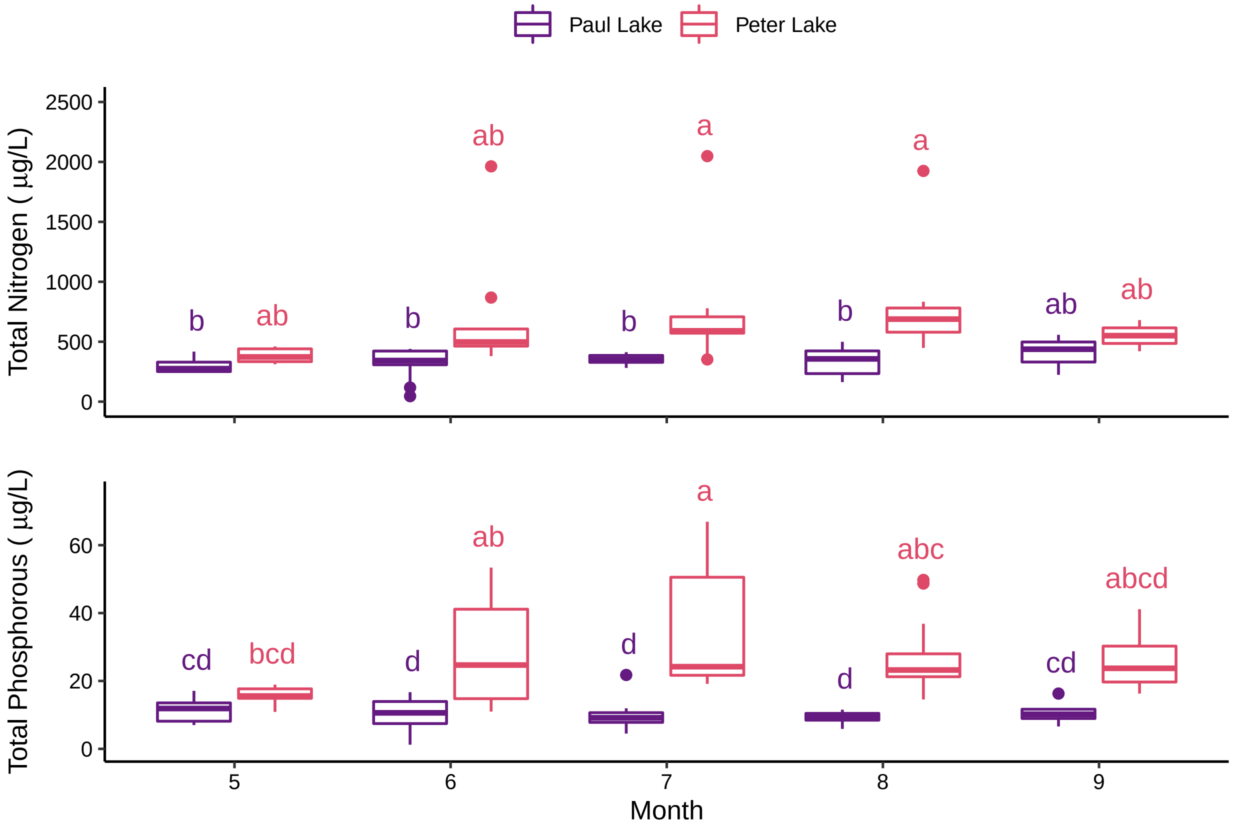
```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).

## Warning: Removed 1 rows containing non-finite values (stat_summary).
```



```
#8 Create Cowplot with common legend and common x-label (on the bottom)
ppPlotGrid <- plot_grid(tn.anova.plot + xlab("") + theme(axis.text.x = element_blank()),
                        tp.anova.plot + theme(legend.position = "none"),
                        align = "v", ncol = 1, rel_heights = c(2,1.5))
```

```
## Warning: Removed 23 rows containing non-finite values (stat_boxplot).

## Warning: Removed 23 rows containing non-finite values (stat_summary).

## Warning: Removed 1 rows containing non-finite values (stat_boxplot).

## Warning: Removed 1 rows containing non-finite values (stat_summary).
```

```
print(ppPlotGrid)
```


Paul Lake
Peter Lake
Total Nitrogen (µg/L)
2500
2000
1500
1000
500
0
b
ab
b
ab
b
a
b
a
ab
ab
Total Phosphorous (µg/L)
60
40
20
0
cd
bcd
d
ab
d
a
d
abc
cd
abcd
5
6
7
8
9
Month