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

Pierre Mishra

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
getwd()
```

```
## [1] "C:/Users/Peaceful Pierre/Documents/Academics/Spring 2020/Environmental Data Analytics/Environme
```

```
library("tidyverse")
library("cowplot")
library("agricolae")
peterpaul <- read.csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv")

#2
peterpaul$sampledate <- as.Date(peterpaul$sampledate, format = "%Y-%m-%d")
head(peterpaul$sampledate)
```

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

```r
class(peterpaul$sampledate)
```

```
## [1] "Date"
```

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.

```r
class(peterpaul$year4)
```

```
## [1] "integer"
```

```r
peterpaul_surface <- filter(peterpaul, depth == 0.00 &
                              (year4 == 1993 |
                                 year4 == 1994|
                                 year4 == 1995|
                                 year4 == 1996))

peterpaul_surface$month <- as.factor(peterpaul_surface$month)
class(peterpaul_surface$month)
```

```
## [1] "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 because here I have a continuous response variable and two categorical explanatory variables and I am also interested in the interection effects between the explanatory variables (months and lakes). I want to see if total phosphorus or total nitrogen (continuous response) is higher in Peter or Paul Lake (first categorical variable) and see if these potential differences vary seasonally (second categorical variable).

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.

```r
#5
peterpaul_tn <- aov(data = peterpaul_surface, tn_ug ~ lakename * month)
summary (peterpaul_tn) # no significant interaction effects, but significant main effect of lake
```

```
##                Df  Sum Sq Mean Sq F value   Pr(>F)
## lakename        1 2468595 2468595  36.414 2.91e-08 ***
## month           4  459542  114885   1.695    0.157
## lakename:month  4  288272   72068   1.063    0.379
## Residuals      97 6575834   67792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 23 observations deleted due to missingness
```

```r
grouping_1 <- HSD.test(peterpaul_tn, "lakename", group = TRUE)
grouping_1
```

```
## $statistics
##   MSerror Df     Mean       CV
##   67792.1 97 487.4077 53.41917
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   2         2.806822  0.05
##
## $means
##                tn_ug      std  r     Min      Max      Q25     Q50      Q75
## Paul Lake  336.9293 100.2745 54  45.670  557.812 284.0107 344.243 411.5165
## Peter Lake 640.7253 361.3738 53 312.133 2048.151 448.0490 571.092 692.4860
##
## $comparison
## NULL
##
## $groups
##                tn_ug groups
## Peter Lake 640.7253      a
## Paul Lake  336.9293      b
##
## attr(,"class")
## [1] "group"
```

```r
#6
peterpaul_tp <- aov(data = peterpaul_surface, tp_ug ~ lakename * month)
summary (peterpaul_tp) # significant interaction effects
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## lakename         1  10228   10228  98.914 <2e-16 ***
## month            4    813     203   1.965 0.1043
## lakename:month   4   1014     254   2.452 0.0496 *
## Residuals      119  12305     103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

```r
peterpaul_interaction <- with(peterpaul_surface, interaction (lakename, month))
peterpaul_interaction_anova <- aov(data = peterpaul_surface, tp_ug ~ peterpaul_interaction)
```

```r
grouping_2 <- HSD.test(peterpaul_interaction_anova, "peterpaul_interaction", group = TRUE)
grouping_2
```

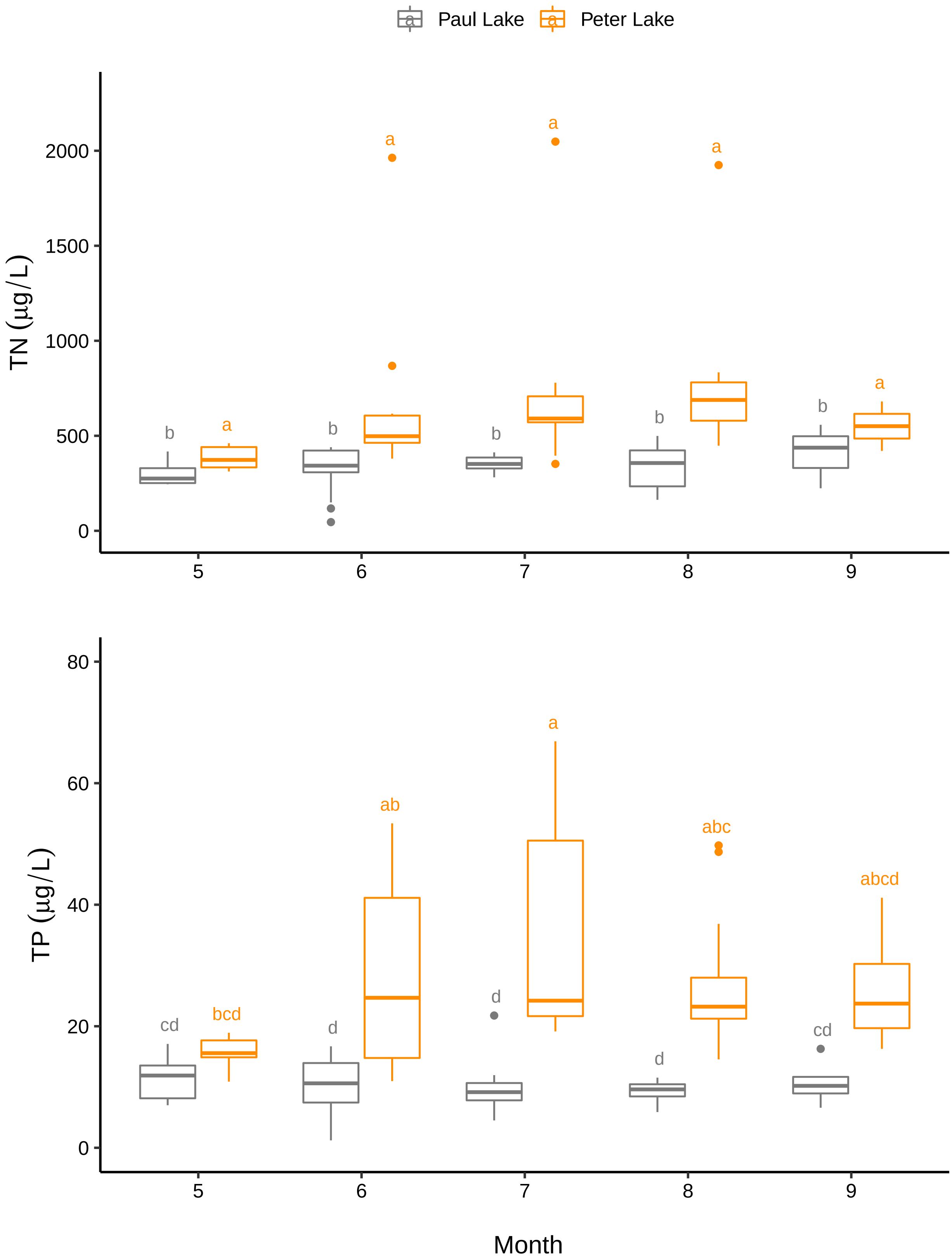
```
## $statistics
##    MSerror  Df     Mean      CV
##   103.4055 119 19.07347 53.3141
##
## $parameters
##    test                name.t ntr StudentizedRange alpha
##   Tukey peterpaul_interaction  10         4.560262  0.05
##
## $means
##                   tp_ug       std  r    Min    Max     Q25     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.

```r
#setting theme
peaceful.theme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")

theme_set(peaceful.theme)

#7
#### total nitrogen
tn_plot <- ggplot(peterpaul_surface, aes(y = tn_ug, x = month, color = lakename)) +
  geom_boxplot() +
  labs(y = expression(TN ~ (mu*g / L)), x = " ", color = " ") +
  theme (legend.position = "top") + ylim (0,2300) +
  stat_summary(geom = "text", fun.y = max, vjust = -1,
               position = position_dodge(.7),  size = 3.5,
               label = c("a", "b", "a", "b", "a", "b",
                         "a", "b", "a", "b")) +
  scale_color_manual(values = c("Paul Lake" = "gray48", "Peter Lake" = "darkorange"))

#### total phosphorus
tp_plot <- ggplot(peterpaul_surface, aes(y = tp_ug, x = month, color = lakename)) +
  geom_boxplot() +
  labs(y = expression(TP ~ (mu*g / L)), x = "\n Month", color = "Lake Names") +
  theme (legend.position = "top") + ylim (0,80) +
  stat_summary(geom = "text", fun.y = max, vjust = -1,
               position = position_dodge(.7),  size = 3.5,
               label = c("bcd", "cd", "ab", "d", "a", "d",
                         "abc", "d", "abcd", "cd")) +
  scale_color_manual(values = c("Paul Lake" = "gray48", "Peter Lake" = "darkorange"))

#8

plot_grid(tn_plot, tp_plot + theme(legend.position="none"),
          nrow = 2, axis = 'lr', align = 'v', rel_heights = c(1,1))
```

```
## 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).
```


Paul Lake
Peter Lake
TN (µg/L)
2000
1500
1000
500
0
a
b
5
6
7
8
9
TP (µg/L)
80
60
40
20
0
cd
bcd
d
ab
a
abc
abcd
Month