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

Kristine Swann

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

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



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

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

```
head(ppp_nut$sampledate)
```

```
## [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.

```r
class(ppp_nut$month)
```

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

```r
ppp_nut_cleaned <-
  ppp_nut %>%
filter(depth == 0) %>%
  filter(year4, c(1993,1994,1995,1996))
ppp_nut_cleaned$month <- factor(ppp_nut_cleaned$month)
write.csv(ppp_nut_cleaned, row.names = FALSE,
          file ="./Data/Processed/ppp_nut_cleaned.csv")
```

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: 2 way anovas with interaction effects; there are multiple categories here - lakename and month, and there may be an interaction between month and lake.

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
shapiro.test(ppp_nut_cleaned$tn_ug[ppp_nut_cleaned$lakename == "Peter Lake"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ppp_nut_cleaned$tn_ug[ppp_nut_cleaned$lakename == "Peter Lake"]
## W = 0.71488, p-value = 1.676e-12
```

```r
shapiro.test(ppp_nut_cleaned$tn_ug[ppp_nut_cleaned$lakename == "Paul Lake"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ppp_nut_cleaned$tn_ug[ppp_nut_cleaned$lakename == "Paul Lake"]
## W = 0.98798, p-value = 0.514
```

```r
shapiro.test(ppp_nut_cleaned$tn_ug[ppp_nut_cleaned$month == "5"])
```

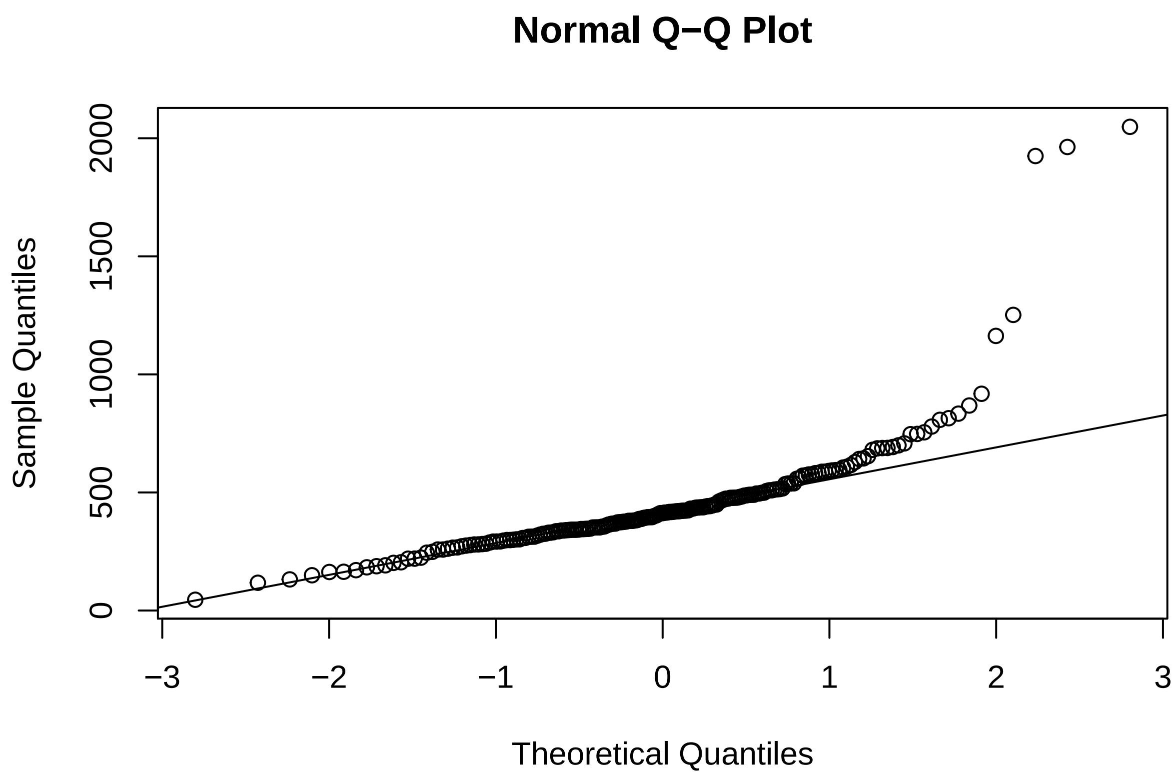
```
##
##  Shapiro-Wilk normality test
##
## data:  ppp_nut_cleaned$tn_ug[ppp_nut_cleaned$month == "5"]
## W = 0.94485, p-value = 0.1605
```

```r
shapiro.test(ppp_nut_cleaned$tn_ug[ppp_nut_cleaned$month == "6"])
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  ppp_nut_cleaned$tn_ug[ppp_nut_cleaned$month == "6"]
## W = 0.69654, p-value = 2.278e-09
```

```r
#some normal, some not normal

qqnorm(ppp_nut_cleaned$tn_ug); qqline(ppp_nut_cleaned$tn_ug) #right tailed
```




```r
# Test for equal variance
bartlett.test(ppp_nut_cleaned$tn_ug ~ ppp_nut_cleaned$lakename)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  ppp_nut_cleaned$tn_ug by ppp_nut_cleaned$lakename
## Bartlett's K-squared = 98.863, df = 1, p-value < 2.2e-16
```

```r
bartlett.test(ppp_nut_cleaned$tn_ug ~ ppp_nut_cleaned$month)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  ppp_nut_cleaned$tn_ug by ppp_nut_cleaned$month
## Bartlett's K-squared = 32.856, df = 4, p-value = 1.278e-06
```

```r
#both sig variance

pp.tn.anova.2way <- aov(data = ppp_nut_cleaned, tn_ug ~ lakename * month)
summary(pp.tn.anova.2way)
```

```
##                  Df   Sum Sq Mean Sq F value   Pr(>F)    
## lakename          1  1652994 1652994  29.002 2.15e-07 ***
## month             4   273647   68412   1.200    0.312    
## lakename:month    4   181160   45290   0.795    0.530    
## Residuals       187 10658205   56996                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 67 observations deleted due to missingness
```

```
#Month was not significant on its own and the interaction between month and lakename wasn't signficant

tn.interaction <- with(ppp_nut_cleaned, interaction(lakename, month))
pp.tn.anova.2way2 <- aov(data = ppp_nut_cleaned, tn_ug ~ tn.interaction)

tn.groups <- HSD.test(pp.tn.anova.2way2, "tn.interaction", group = TRUE)
tn.groups
```

```
## $statistics
##    MSerror  Df     Mean       CV
##   56995.75 187 456.2056 52.33119
## 
## $parameters
##    test         name.t ntr StudentizedRange alpha
##   Tukey tn.interaction  10         4.528779  0.05
## 
## $means
##                  tn_ug       std  r     Min      Max      Q25      Q50      Q75
## Paul Lake.5  356.2339  98.64852 13 244.870  538.000 279.6200 340.2520 417.3450
## Paul Lake.6  362.9227 131.88118 28  45.670  628.625 305.7615 382.8905 426.7722
## Paul Lake.7  369.2854  66.47625 27 191.370  506.000 335.6635 355.7950 407.8055
## Paul Lake.8  372.5497 108.61774 25 163.148  576.302 312.8900 383.0000 431.4470
## Paul Lake.9  344.0663 166.40933  6 164.080  557.812 209.1823 330.5980 467.0992
## Peter Lake.5 405.0493  97.93447 14 272.000  593.138 343.2875 363.9700 456.7887
## Peter Lake.6 554.6129 324.99144 27 219.720 1962.902 401.3825 509.2440 596.5010
## Peter Lake.7 571.6838 368.75258 26 131.830 2048.151 355.0007 514.2170 664.6598
## Peter Lake.8 612.4790 357.72223 26 201.770 1924.631 369.1100 552.5425 733.9880
## Peter Lake.9 459.4612 128.39261  5 345.000  680.558 417.1900 420.3780 434.1800
## 
## $comparison
## NULL
## 
## $groups
##                  tn_ug groups
## Peter Lake.8 612.4790      a
## Peter Lake.7 571.6838     ab
## Peter Lake.6 554.6129    abc
## Peter Lake.9 459.4612    abc
## Peter Lake.5 405.0493    abc
## Paul Lake.8  372.5497     bc
## Paul Lake.7  369.2854     bc
## Paul Lake.6  362.9227      c
## Paul Lake.5  356.2339      c
## Paul Lake.9  344.0663      c
## 
## attr(,"class")
```

```
## [1] "group"
```

```r
#The groupings here show that there's a lot of overap in significance between months and lakes - ie the

#6
shapiro.test(ppp_nut_cleaned$tp_ug[ppp_nut_cleaned$lakename == "Peter Lake"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ppp_nut_cleaned$tp_ug[ppp_nut_cleaned$lakename == "Peter Lake"]
## W = 0.87354, p-value = 3.141e-09
```

```r
shapiro.test(ppp_nut_cleaned$tp_ug[ppp_nut_cleaned$lakename == "Paul Lake"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ppp_nut_cleaned$tp_ug[ppp_nut_cleaned$lakename == "Paul Lake"]
## W = 0.85097, p-value = 3.596e-10
```

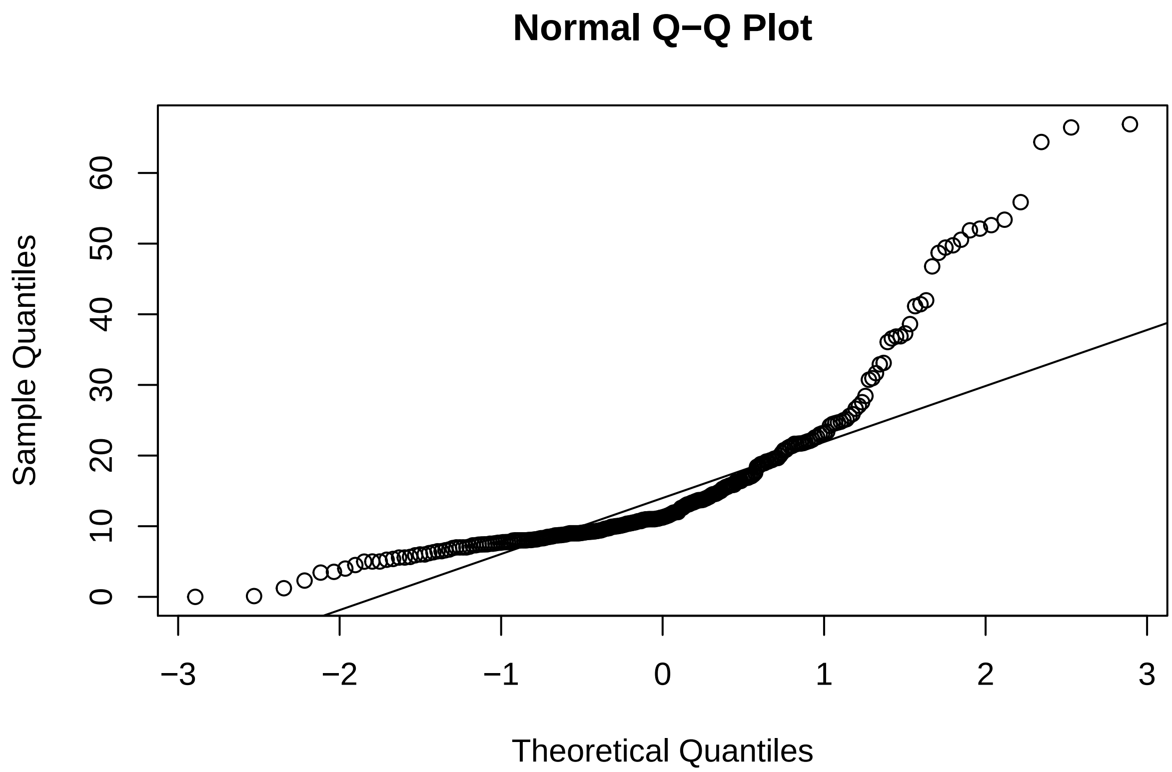
```r
shapiro.test(ppp_nut_cleaned$tp_ug[ppp_nut_cleaned$month == "5"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ppp_nut_cleaned$tp_ug[ppp_nut_cleaned$month == "5"]
## W = 0.94993, p-value = 0.2135
```

```r
shapiro.test(ppp_nut_cleaned$tp_ug[ppp_nut_cleaned$month == "6"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ppp_nut_cleaned$tp_ug[ppp_nut_cleaned$month == "6"]
## W = 0.78439, p-value = 6.675e-09
```

```r
#Both lakes are not normally distributed for TP, the months vary in whether their distributions are nor

qqnorm(ppp_nut_cleaned$tp_ug); qqline(ppp_nut_cleaned$tp_ug) #funky and super NOT normal distribution -
```




```r
bartlett.test(ppp_nut_cleaned$tp_ug ~ ppp_nut_cleaned$lakename)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  ppp_nut_cleaned$tp_ug by ppp_nut_cleaned$lakename
## Bartlett's K-squared = 141.45, df = 1, p-value < 2.2e-16
```

```r
bartlett.test(ppp_nut_cleaned$tp_ug ~ ppp_nut_cleaned$month)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  ppp_nut_cleaned$tp_ug by ppp_nut_cleaned$month
## Bartlett's K-squared = 29.929, df = 4, p-value = 5.06e-06
```

```r
#both sig variance

pp.tp.anova.2way <- aov(data = ppp_nut_cleaned, tp_ug ~ lakename * month)
summary(pp.tp.anova.2way)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## lakename          1   8046    8046  74.199 7.72e-16 ***
## month             4    399     100   0.919    0.453
## lakename:month    4    892     223   2.056    0.087 .
## Residuals       253  27435     108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## 1 observation deleted due to missingness
```

```r
#Month was not significant on its own and the interaction between month and lakename wasn't signficant

tp.interaction <- with(ppp_nut_cleaned, interaction(lakename, month))
pp.tp.anova.2way2 <- aov(data = ppp_nut_cleaned, tp_ug ~ tp.interaction)

tp.groups <- HSD.test(pp.tp.anova.2way2, "tp.interaction", group = TRUE)
tp.groups
```

```
## $statistics
##    MSerror  Df     Mean       CV
##   108.4382 253 15.82452 65.80524
##
## $parameters
##    test         name.t ntr StudentizedRange alpha
##   Tukey tp.interaction  10         4.514467  0.05
##
## $means
##                  tp_ug       std  r   Min    Max      Q25    Q50      Q75
## Paul Lake.5  12.63408  5.828445 13 7.001 25.000  8.00000 10.697 13.68900
## Paul Lake.6  10.13697  4.365819 36 0.110 17.557  7.26925 10.356 13.46475
## Paul Lake.7  10.17403  5.788561 37 2.305 36.070  7.75300  9.000 10.72800
## Paul Lake.8   9.51425  1.777077 36 5.879 13.873  8.04475  9.561 10.62600
## Paul Lake.9  10.83878  4.360943  9 6.592 19.370  7.41900 10.080 11.67100
## Peter Lake.5 13.95943  4.544036 14 5.650 23.000 11.03500 13.919 16.50375
## Peter Lake.6 19.90478 14.574321 36 0.000 53.388  9.55050 15.580 24.63200
## Peter Lake.7 24.20532 16.838706 37 5.000 66.893 11.23000 21.664 27.05600
## Peter Lake.8 22.33789 11.840371 37 6.190 49.757 13.22200 21.112 27.55400
## Peter Lake.9 22.75900 16.621186  8 6.000 52.615  9.30000 18.550 30.26025
##
## $comparison
## NULL
##
## $groups
##                 tp_ug groups
## Peter Lake.7 24.20532      a
## Peter Lake.9 22.75900     ab
## Peter Lake.8 22.33789     ab
## Peter Lake.6 19.90478     ab
## Peter Lake.5 13.95943    abc
## Paul Lake.5  12.63408     bc
## Paul Lake.9  10.83878     bc
## Paul Lake.7  10.17403     bc
## Paul Lake.6  10.13697     bc
## Paul Lake.8   9.51425      c
##
## attr(,"class")
## [1] "group"
```

```r
#The groupings here are similar to those for TN. They show that there's a lot of overap in significance

#NOTE: the HSD tests are superfluous because the interaction didn't exist. Really, when groupings are d
```

7. Create two plots, with TN (plot 1) or TP (plot 2) as the response variable and month and lake as the

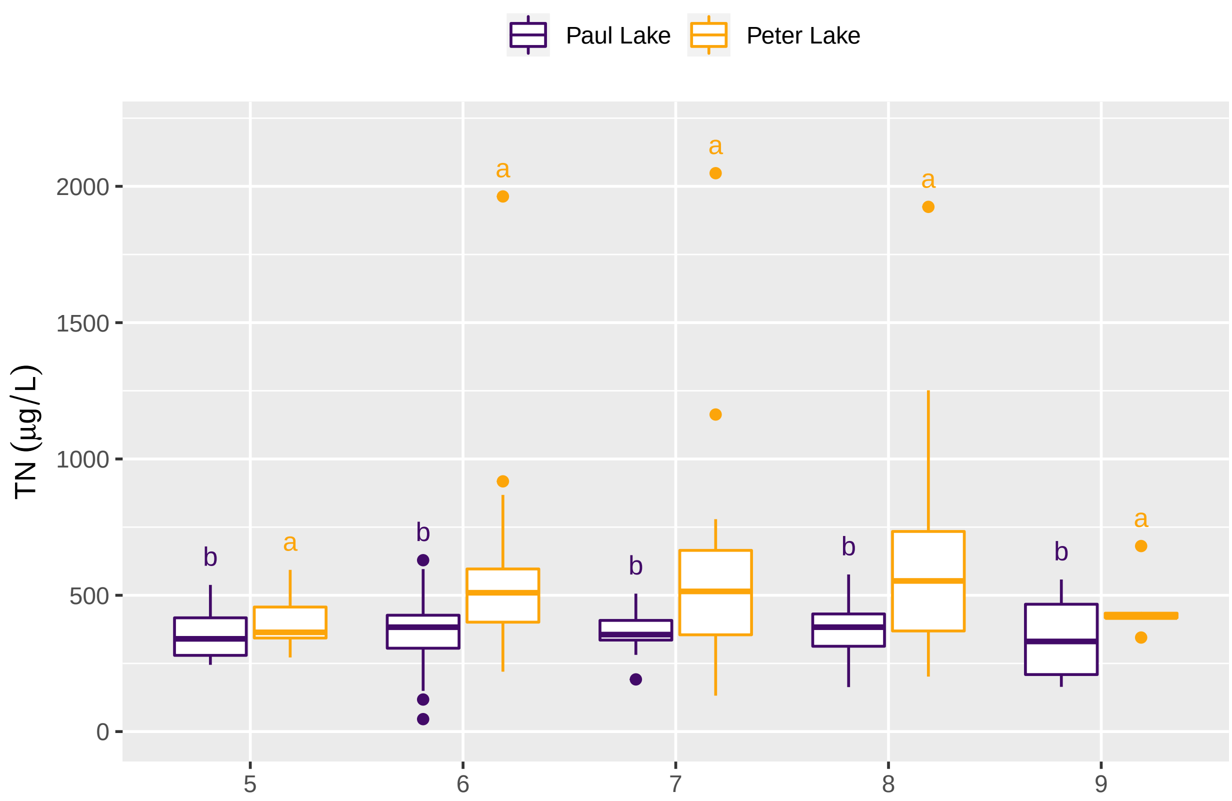
predictor variables. Hint: you may use some of the code you used for your visualization assignment. Assign groupings with letters, as determined from your tests. Adjust your axes, aesthetics, and color palettes in accordance with best data visualization practices.

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
tn.anova.plot <- ggplot(ppp_nut_cleaned, aes(y = tn_ug, x = month, color =  lakename)) +
  geom_boxplot()+
stat_summary(geom = "text", fun.y = max, vjust = -1, size = 3.5,
             label = c("a", "b", "a", "b","a","b","a","b","a", "b"), position = position_dodge(0.75), sl
  theme(legend.position= "top")+
  scale_color_viridis_d(option = "inferno", begin = 0.2, end = 0.8)+
  labs(x = " ", y = expression(TN ~ (mu*g / L)), color= "")+
    ylim(0, 2200)
print(tn.anova.plot)
```

```
## Warning: Removed 67 rows containing non-finite values (stat_boxplot).
```

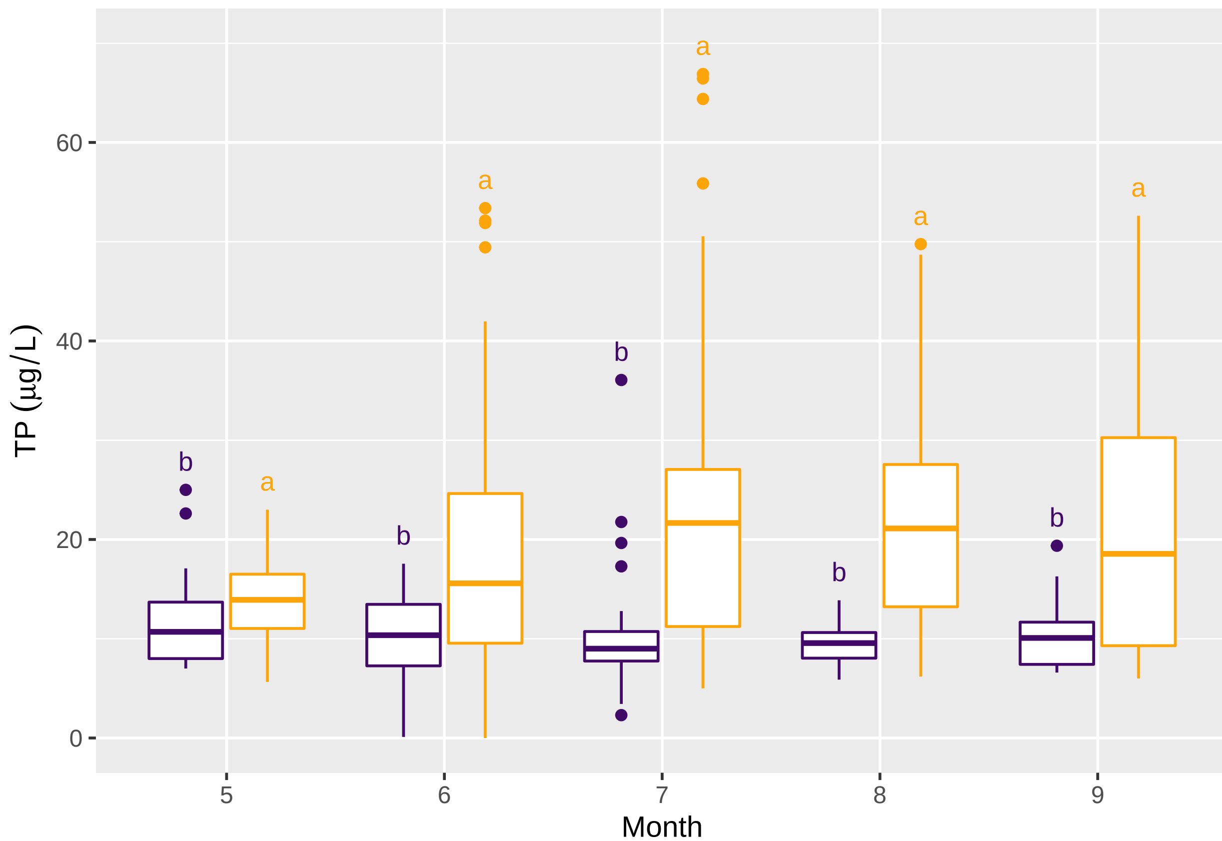
```
## Warning: Removed 67 rows containing non-finite values (stat_summary).
```




```
##$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$~~~~~~~~~~~~~~~~~~~~~~~~~~$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$

tp.anova.plot <- ggplot(ppp_nut_cleaned, aes(y = tp_ug, x = month, color =  lakename)) +
  geom_boxplot()+
stat_summary(geom = "text", fun.y = max, vjust = -1, size = 3.5,
```

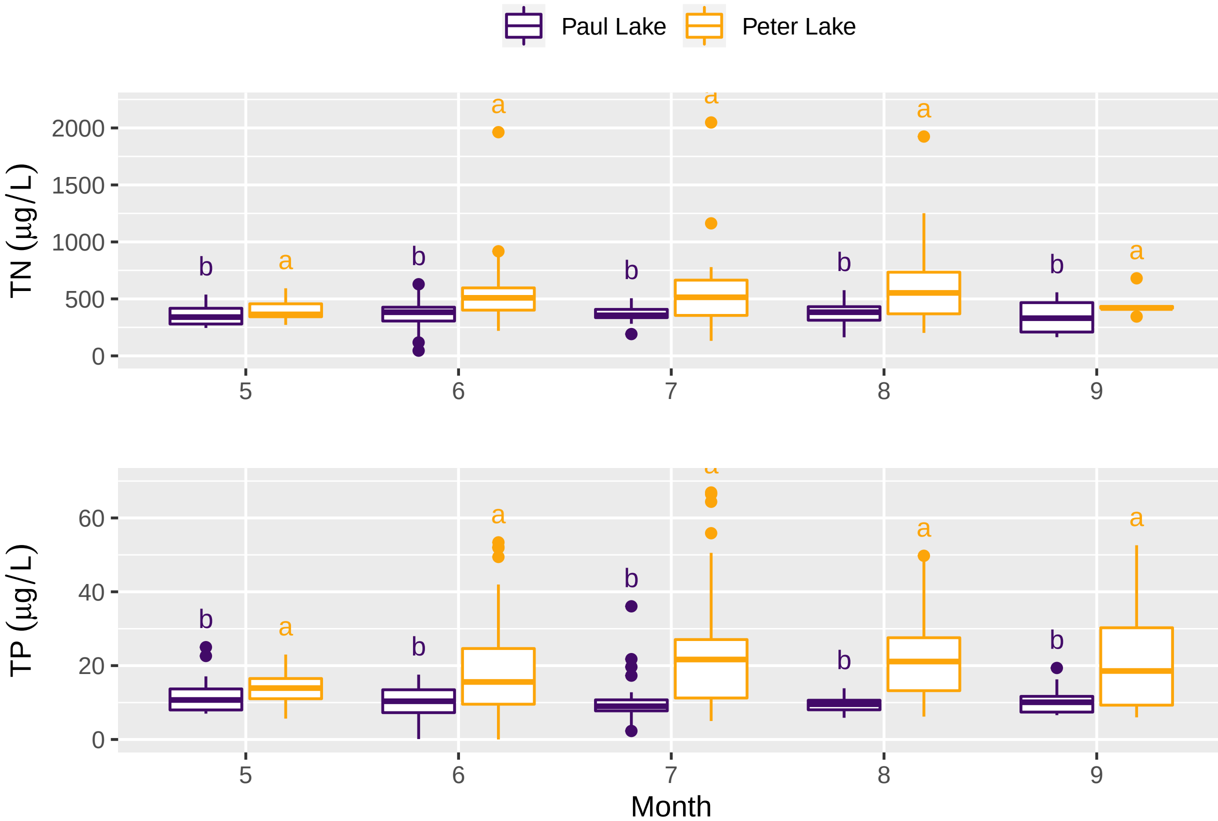
```r
             label = c("a", "b", "a", "b","a","b","a","b","a", "b"), position = position_dodge(0.75))+
  scale_color_viridis_d(option = "inferno", begin = 0.2, end = 0.8)+
  theme(legend.position = "none")+
  labs(x = "Month", y = expression(TP ~ (mu*g / L)))+
  ylim(0, 70)

print(tp.anova.plot)
```

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

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




```r
#8
plot_grid(tn.anova.plot,tp.anova.plot, nrow=2, align = "v", rel_heights = c(1.25, 1))
```

```
## Warning: Removed 67 rows containing non-finite values (stat_boxplot).

## Warning: Removed 67 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
TP (µg/L)
60
40
20
0
5
6
7
8
9
Month
a
b